Argument & Computation -1 (2024) 1–41
DOI 10.3233/AAC-230008
IOS Press

Cross-genre argument mining: Can language models automatically fill in missing discourse markers?

Gil Rocha [a,*], Henrique Lopes Cardoso [a], Jonas Belouadi [b] and Steffen Eger [b]
[a] *Laboratório de Inteligência Artificial e Ciência de Computadores (LIACC), Faculdade de Engenharia da Universidade do Porto, Porto, Portugal*
E-mails: gil.rocha@fe.up.pt, hlc@fe.up.pt
[b] *Natural Language Learning Group (NLLG), Bielefeld University & University of Mannheim, Germany*
E-mails: jonas.belouadi@uni-bielefeld.de, steffen.eger@uni-mannheim.de

Abstract. Available corpora for Argument Mining differ along several axes, and one of the key differences is the presence (or absence) of discourse markers to signal argumentative content. Exploring effective ways to use discourse markers has received wide attention in various discourse parsing tasks, from which it is well-known that discourse markers are strong indicators of discourse relations. To improve the robustness of Argument Mining systems across different genres, we propose to automatically augment a given text with discourse markers such that all relations are explicitly signaled. Our analysis unveils that popular language models taken out-of-the-box fail on this task; however, when fine-tuned on a new heterogeneous dataset that we construct (including synthetic and real examples), they perform considerably better. We demonstrate the impact of our approach on an Argument Mining downstream task, evaluated on different corpora, showing that language models can be trained to automatically fill in discourse markers across different corpora, improving the performance of a downstream model in some, but not all, cases. Our proposed approach can further be employed as an assistant tool for better discourse understanding.

Keywords: Computational linguistics, machine learning, language models, discourse analysis, argument mining

1. Introduction

Argument Mining (ArgMining) is a discourse parsing task that aims to automatically extract structured arguments from text. In general, an argument in NLP and machine learning is a graph-based structure, where nodes correspond to Argumentative Discourse Units (ADUs), which are connected via argumentative relations (e.g., support or attack) [36]. Available corpora for ArgMining differ along several axes, such as language, genre, domain, and annotation schema [33,36]. One key aspect that differs across different corpora (and even across different articles in the same corpus) is the presence (or absence) of discourse markers (DMs) [18,57]. These DMs are lexical clues that often precede ADUs [2,30,45,46,56,62,63]. DMs have been studied for several years under many different names [21], including discourse connectives, discourse particles, pragmatic markers, and cue phrases. We adopt the definition of discourse markers proposed by Fraser [20]: DMs are lexical expressions (“and”, “but”,

*Corresponding author. E-mail: gil.rocha@fe.up.pt.

"because", "however", "as a conclusion", etc.) used to relate discourse units. With certain exceptions, they signal a relationship between the interpretation of the discourse unit they introduce (S2) and the prior discourse unit (S1). DMs might belong to different syntactic classes, drawn primarily from the syntactic classes of conjunctions, adverbials, and prepositional phrases.

Exploring effective ways to use DMs has received wide attention in various NLP tasks [47,61], including ArgMining related tasks [30,46]. In discourse parsing [38,51], DMs are known to be strong cues for the identification of discourse relations [7,39]. Similarly, for ArgMining, the presence of DMs are strong indicators for the identification of ADUs [18] and for the overall argument structure [30,63] (e.g., some DMs are clear indicators of the ADU role, such as premise, conclusion, or major claim).

The absence of DMs makes the task more challenging, requiring the system to more deeply capture semantic relations between ADUs [40].

To close the gap between these two scenarios (i.e., relations explicitly signaled with DMs vs. implicit relations), we ask whether recent large language models (LLMs) such as BART [35], T5 [54] and ChatGPT,[1] can be used to automatically augment a given text with DMs such that all relations are explicitly signaled. Due to the impressive language understanding and generation abilities of recent LLMs, we speculate that such capabilities could be leveraged to automatically augment a given text with DMs. However, our analysis unveils that such language models (LMs), when employed in a zero-shot setting, underperform in this task. To overcome this challenge, we hypothesize that a sequence-to-sequence (Seq2Seq) model fine-tuned to augment DMs in an end-to-end setting (from an original to a DM-augmented text) should be able to add coherent and grammatically correct DMs, thus adding crucial signals for ArgMining systems.

Our second hypothesis is that downstream ArgMining models can profit from automatically added DMs because these contain highly relevant signals for solving ArgMining tasks, such as ADU identification and classification. Moreover, given that the Seq2Seq model is fine-tuned on heterogeneous data, we expect it to perform well across different genres. To this end, we demonstrate the impact of our approach on an ArgMining downstream task, evaluated on different corpora.

Our experiments indicate that the proposed Seq2Seq models can augment a given text with relevant DMs; however, the lack of consistency and variability of augmented DMs impacts the capability of downstream task models to systematically improve the scores across different corpora. Overall, we show that the DM augmentation approach improves the performance of ArgMining systems in some corpora and that it provides a viable means to inject explicit indicators when the argumentative reasoning steps are implicit. Besides improving the performance of ArgMining systems, especially across different genres, we believe that this approach can be useful as an assistant tool for discourse understanding to improve the readability of the text and the transparency of argument exposition, e.g., in education contexts.

In summary, our main contributions are: (i) we propose a synthetic template-based test suite, accompanied by automatic evaluation metrics and an annotation study, to assess the capabilities of state-of-the-art LMs to predict DMs; (ii) we analyze the capabilities of LMs to augment text with DMs, finding that they underperform in this task; (iii) we compile a heterogeneous collection of DM-related datasets on which we fine-tune LMs, showing that we can substantially improve their ability for DM augmentation, and (iv) we evaluate the impact of end-to-end DM augmentation in a downstream task and find that it improves the performance of ArgMining systems in some, but not all, cases. The code for replicating our experiments is publicly available.[2]

[1] https://openai.com/blog/chatgpt/

[2] https://github.com/GilRocha/dm-augment-xgenre-argmining

2. Discourse marker augmentation approach – high-level description

This section provides a high-level description of the proposed DM augmentation approach.

We start our analysis by assessing whether recent LMs can predict grammatically and coherent DMs (Section 3). To this end, we set up a "fill-in-the-mask" DM prediction task, based on a challenging synthetic testbed proposed in this work (Section 3.1), that allows us to evaluate whether language models are sensitive to specific semantically-critical edits in the text. Furthermore, in Section 3.2, we also propose automatic evaluation metrics (validated with an annotation study, Section 3.6) that can be employed to evaluate the quality of the DM predictions provided by the models.

However, the ultimate goal of this work is to automatically augment a text with DMs in an end-to-end setting (i.e., from raw text to a DM-augmented text). This task is framed as a Seq2Seq problem (Section 5), where end-to-end DM augmentation models must automatically (a) identify where DMs should be added and (b) determine which DMs should be added based on the context. To assess the impact of the proposed DM augmentation approach in a downstream task, we study the ArgMining task of ADU identification and classification (Section 6). Section 4 introduces the corpora used in the end-to-end DM augmentation (Section 5) and downstream task (Section 6) experiments.

Figure 1 illustrates how the proposed DM augmentation approach differs from conventional approaches to address the downstream task studied in this paper (ADU identification and classification).

At the top of Fig. 1, we illustrate the conventional approach to address the ArgMining downstream task. "Original Text" was extracted from the Persuasive Essays corpus (PEC) [63] and contains explicit DMs provided by the essay's author (highlighted in underlined text). The downstream task model re-

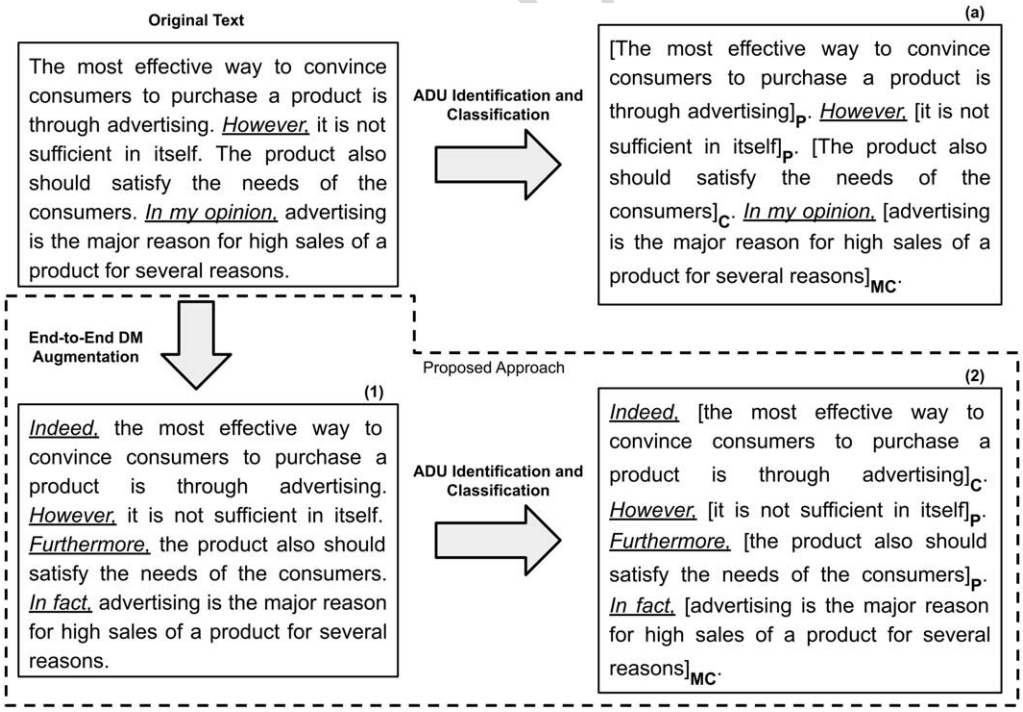


Fig. 1. Illustration of proposed DM augmentation approach. We include annotations for ADU boundaries (square brackets) and labels (acronyms in subscript, see Section 4 for details regarding ADU labels). The underlined text highlights the presence of DMs.

ceives as input the "Original Text", and the goal is to identify and classify the ADUs (according to a corpus-specific set of ADU labels, such as Premise (P), Claim (C), Major Claim (MC), etc.). This downstream task is formulated as a token-level sequence tagging problem. We illustrate the output of a downstream task model in the box (a) from Fig. 1 as follows: ADU boundaries (square brackets) and labels (acronyms in subscript, see Section 4 for details regarding ADU labels). The gold annotation for this example is: *"[The most effective way to convince consumers to purchase a product is through advertising]$_{C}$. However, [it is not sufficient in itself]$_{P}$. [The product also should satisfy the needs of the consumers]$_{P}$. In my opinion, [advertising is the major reason for high sales of a product for several reasons]$_{MC}$."*. Comparing the prediction in box (a) with the gold annotation, we can observe that the downstream task model identifies the ADUs and corresponding boundaries correctly; however, in terms of classification, the model labels the first and third ADUs incorrectly.

In the approach proposed in this paper, in the first step, we employ DM augmentation models to augment the text with DMs in an end-to-end setting (as detailed in Section 5). As illustrated in box (1) from Fig. 1, the output from this initial step is a text augmented with DMs, containing useful indicators to unveil the argumentative content and improve the readability of the text. Then, in a second step, to assess the impact of the proposed end-to-end DM augmentation approach in a downstream task, we evaluate whether downstream task models can take profit from the DMs automatically added to the text (Section 6), as illustrated in box (2) from Fig. 1. Comparing the prediction in box (b) with the gold annotation, we can observe that the downstream task model performs both tasks (i.e., ADU identification and classification) correctly, taking advantage of the useful indicators provided by the DM augmentation model.

3. Fill-in-the-mask discourse marker prediction

We now assess whether current state-of-the-art language models can predict coherent and grammatically correct DMs. To this end, we create the *Template-based Arguments for Discourse Marker prediction (TADM)* dataset that allows us to evaluate whether language models are sensitive to specific semantically-critical edits in the text. These targeted edits are based on the DMs that precede each ADU and the keyword that defines the claim's stance (the positioning of the claim towards a given topic, e.g., "for" or "against"). When manipulated accordingly, the edits can entail relevant changes in the argument structure.

To illustrate the crucial role of DMs and stance for argument perception, consider Examples 1 and 2 in Fig. 2. In bold, we highlight the stance-revealing word. We show that it is possible to obtain different text sequences with opposite stances, illustrating the impact that a single word (the stance in this case) can have on the structure of the argument. Indeed, the role of the premises changes according to the stance (e.g., "X1" is an attacking premise in Example 1 but a supportive premise in Example 2), reinforcing that the semantically-critical edits in the stance impact the argument structure. In underlined text, we highlight the DMs. We remark that by changing the stance, adjustments of the DMs that precede each ADU are required to obtain a coherent text sequence. Furthermore, these examples also show that the presence of explicit DMs improves the readability and makes the argument structure clearer (reducing the cognitive interpretation demands required to unlock the argument structure). In some cases, the presence of explicit DMs is essential (e.g., DMs preceding the attacking premise, such as "Although" and "However" in Examples 1 and 2, respectively) to obtain a sound and comprehensible text. Overall, Fig. 2 illustrates the key property that motivated our TADM dataset: subtle edits in content (i.e., stance

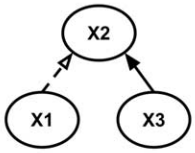
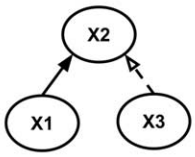
1	"*Although* [ecological concerns add further strain on the economy]$_{X1}$, *I think that* [we should **introduce** carbon taxes]$_{X2}$. *Moreover,* [humanity must embrace clean energy in order to fight climate change]$_{X3}$."	
2	"*Given that* [ecological concerns add further strain on the economy]$_{X1}$, *I think that* [we should **abolish** carbon taxes]$_{X2}$. *However,* [humanity must embrace clean energy in order to fight climate change]$_{X3}$."	

Fig. 2. TADM dataset example. Columns correspond to example id, text sequence, and argument structure (respectively). In the text sequences, we highlight the DMs (underlined) and the stance revealing word (in bold). In the argument structure, nodes correspond to ADUs and arrows to argumentative relations (a solid arrow indicates a "support" relation, a dashed arrow an "attack" relation).

and DMs) might have a relevant impact on the argument structure. On the other hand, some edits do not entail any difference in the argument structure (e.g., the position of ADUs, such as whether the claim occurs before/after the premises).

To assess the sensitivity of language models to capture these targeted edits, we use a "fill-in-the-mask" setup, where some of the DMs are masked, and the models aim to predict the masked content. To assess the robustness of language models, we propose a challenging testbed by providing text sequences that share similar ADUs but might entail different argument structures based on the semantically-critical edits in the text previously mentioned (Section 3.1). To master this task, models are not only required to capture the semantics of the sentence (claim's stance and role of each ADU) but also to take into account the DMs that precede the ADUs (as explicit indicators of other ADUs' roles and positioning).

3.1. Template-based Arguments for Discourse Marker prediction (TADM) dataset

Each sample comprises a claim and one (or two) premise(s) such that we can map each sample to a simple argument structure. Every ADU is preceded by one DM that signals its role. Claims have a clear stance towards a given position (e.g., "we should introduce carbon taxes"). This stance is identifiable by a keyword in the sentence (e.g., "introduce"). To make the dataset challenging, we also provide samples with opposite stances (e.g., "introduce" vs. "abolish"). We have one premise in support of a given $\langle claim, stance \rangle$, and another against it. For the opposite stance, the roles of the premises are inverted (i.e., the supportive premise for the claim with the original stance becomes the attacking premise for the claim with the opposite stance). For the example in Fig. 2, we have the following core elements: claim = "we should <STANCE> carbon taxes", original stance = "introduce", opposite stance = "abolish", premise support (original stance) = "humanity must embrace clean energy in order to fight climate change", and premise attack (original stance) = "ecological concerns add further strain on the economy".

Based on these core elements, we follow a template-based procedure to generate different samples. The templates are based on a set of configurable parameters: number of ADUs $\in \{2, 3\}$; stance role $\in \{original, opposite\}$; claim position $\in \{1, 2\}$; premise role $\in \{support, attack\}$; supportive premise position $\in \{1, 2\}$; prediction type $\in \{dm_1, dm_2, dm_3\}$. Additional details regarding the templates can be

found in Appendix A. We generate one sample for each possible configuration, resulting in 40 samples generated for a given instantiation of the aforementioned core elements.

DMs are added based on the role of the ADU that they precede, using a fixed set of DMs (details provided in Appendix B). Even though using a fixed set of DMs comes with a lack of diversity in our examples, the main goal is to add gold DMs consistent with the ADU role they precede. We expect that the language models will naturally add some variability in the set of DMs predicted, which we compare with our gold DMs set (using either lexical and semantic-level metrics, as introduced in Section 3.2).

Examples 1 and 2 in Fig. 2 illustrate two samples from the dataset. In these examples, all possible masked tokens are revealed (i.e., the underlined DMs). The parameter "prediction type" will dictate which of these tokens will be masked for a given sample.

To instantiate the "claim", "original stance", "premise support" and "premise attack", we employ the Class of Principled Arguments (CoPAs) set provided by Bilu et al. [4]. CoPAs are sets of propositions that are often used when debating a recurring theme. Associated with a recurring theme, Bilu et al. [4] provide a set of motions that are relevant to the recurring theme. In the context of a debate, a motion is a proposal that is to be deliberated by two sides. For example, for the theme "Clean energy", we can find the motion phrasing "we should introduce carbon taxes". We use these motions as claims in our TADM dataset. For each CoPA, Bilu et al. [4] provide two propositions that people tend to agree as supporting different points of view for a given theme. We use these propositions as supportive and attacking premises towards a given claim. Further details regarding this procedure can be found in Appendix C.

The test set contains 15 instantiations of the core elements (all from different CoPAs), resulting in a total of 600 samples (40 samples per instantiation $\times$ 15 instantiations). For the train and dev set, we use only the original stance role (i.e., stance role $=$ *original*), reducing to 20 the number of samples generated for each core element instantiation. We have 251 and 38 instantiations of the core elements[3] resulting in a total of 5020 and 760 samples (for the train and dev set, respectively). To avoid overlap of the core elements across different partitions, a CoPA being used in one partition is not used in another.

3.2. Automatic evaluation metrics

To evaluate the quality of the predictions provided by the models (compared to the gold DMs), we explore the following metrics. (1) *Embeddings-based text similarity*: "word embs" based on pre-trained embeddings from spaCy library,[4] "retrofit embs" based on pre-trained embeddings from LEAR [68], "sbert embs" using pre-trained sentence embeddings from the SBERT library [55]. (2) *Argument marker sense* ("arg marker"): DMs are mapped to argument marker senses (i.e., "forward", "backward", "thesis", and "rebuttal") [63]; we consider a prediction correct if the predicted and gold DM senses match. (3) *Discourse relation sense* ("disc rel"): we proceed similar to "arg marker" but using discourse relation senses [12]. These senses are organized hierarchically into 3 levels (e.g., "Contingency.Cause.Result") – in this work, we consider only the first level of the senses (i.e., "Comparison", "Contingency", "Expansion", and "Temporal"). Appendix D provides additional details on how these metrics are computed.

For all metrics, scores range from 0 to 1, and obtaining higher scores means performing better (on the corresponding metric).

[3]The train set contains 12 CoPAs, and the dev set contains 2. To increase the number of instantiations in these partitions, we include all possible motions for each CoPA. Given that each CoPA is associated with a varied number of motions, this means that the total number of instantiations (which are based on the motions) differs for each CoPA. For this reason, we indicate the total number of instantiations for each partition. Note that different instantiations based on the same CoPA have different claim content but share the same premises (i.e., premises are attached to a given CoPA).

[4]https://spacy.io/

3.3. Models

We employ the following LMs: BERT ("bert-base-cased") [15], XLM-RoBERTa (XLM-R) ("xlm-roberta-base") [11], and BART ("facebook/bart-base") [35]. As a first step in our analysis, we frame the DM augmentation problem as a single mask token prediction task.

For BART, we also report results following a Seq2Seq setting (`BART V2`), where the model receives the same input sequence as the previously mentioned models (text with a single mask token) and returns a text sequence as output. BART is a Seq2Seq model which uses both an encoder and a decoder from a Transformer-based architecture [67]. Consequently, we can explore the Seq2Seq capabilities of this model to perform end-to-end DM augmentation, contrary to the other models which only contain the encoder component (hence, limited to the single mask-filling setting). Including `BART V2` in this analysis provides a means to compare the performance between these formulations: simpler single-mask prediction from `BART V1` vs Seq2Seq prediction from `BART V2`. For the Seq2Seq setting, determining the predicted DM is more laborious, requiring a comparison between the input and output sequence. Based on a diff-tool implementation,[5] we determine the subsequence from the output sequence (in this case, we might have a varied number of tokens being predicted, from zero to several tokens) that matches the "<mask>" token in the input sequence. Note that the output sequence might contain further differences as compared to the input sequence (i.e., the model might perform further edits to the sequence besides mask-filling); however, these differences are discarded for the fill-in-the-mask DM prediction task evaluation. As detailed in Appendix E, in a few cases, we observed minor additional edits being performed that do not change the overall meaning of the content, mostly related to specific grammatical edits.

3.4. Experiments

Table 1 shows the results obtained on the test set of the TADM dataset for the fill-in-the-mask DM prediction task.

Table 1

Fill-in-the-mask DM predictions results. Evaluated on the test set of the TADM dataset. Columns: automatic evaluation metrics (Section 3.2). "Zero-shot": results obtained by the LMs described in Section 3.3 when employed in a zero-shot setting. "Fine-tuned (BERT)": fine-tuned BERT on the TADM dataset (including the different settings explored for fine-tuning experiments, detailed in Section 3.4)

	word embs	retrofit embs	sbert embs	arg marker	disc rel
Zero-shot					
BERT	**0.7567**	**0.6336**	**0.4605**	0.2444	**0.4361**
XLM-R	0.7196	0.6080	0.4460	**0.2500**	0.3944
BART V1	0.6225	0.4599	0.3165	0.1861	0.3194
BART V2	0.6934	0.4325	0.3093	0.0833	0.2417
Fine-tuned (BERT)					
2 ADUs	0.7382	0.5981	0.5659	0.4528	0.6083
3 ADUs	0.9687	0.9296	0.8688	0.8333	0.8333
pred type 1	0.6886	0.6024	0.5879	0.5278	0.5417
pred type 2	0.6935	0.6080	0.5944	0.5361	0.5361
all	**0.9786**	**0.9519**	**0.9103**	**0.8861**	**0.8861**

[5] https://docs.python.org/3/library/difflib.html

In a zero-shot setting, BERT performs clearly best, leading all evaluation dimensions except for "arg marker". BART models obtain the lowest scores for all metrics.

We explore fine-tuning the models on the TADM dataset (using the train and dev set partitions). We use BERT for these fine-tuning experiments. Comparing fine-tuned BERT ("all") with zero-shot BERT, we observe clear improvements in all metrics. Increases are particularly striking for "arg marker" and "disc rel" where models improve from 0.24–0.44 to almost 0.9. Thus, we conclude that the DM slot-filling task is very challenging in a zero-shot setting for current LMs, but after fine-tuning, the performance can be clearly improved.

Due to the limitations in terms of diversity of the TADM dataset, we expected some overfitting after fine-tuning the models on this data. To analyze the extent to which models overfit the fine-tuning data, we selected samples from the training set by constraining some of the parameters used to generate the TADM dataset. The test set remains the same as in previous experiments. In BERT ("2 ADUs"), only samples containing a single sentence are included (besides single sentences, the test set also includes samples containing two sentences with 3 ADUs that the model was not trained with). For BERT ("3 ADUs"), only samples containing two sentences with 3 ADUs are included (thus, even though the model is trained to make predictions in the first sentence, it is not explicitly trained to make predictions when only a single sentence is provided, as observed in some samples in the test set). In BERT ("pred type 1"), we only include samples where the mask token is placed at the beginning of the first sentence (consequently, the model is not trained to predict DMs in other positions). For BERT ("pred type 2"), we only include samples where the mask is placed in the DM preceding the second ADU of the first sentence. For both BERT ("pred type 1") and BERT ("pred type 2"), the model is not trained to make predictions in the second sentence; hence, we can analyze whether models generalize well in these cases. Appendix F provides additional details regarding the distribution of samples in the train, dev, and test set according to the discourse-level senses "arg marker" and "disc rel".

Comparing the different settings explored for fine-tuning experiments, we observe that constraining the training data to single-sentence samples ("2 ADUs") or to a specific "pred type" negatively impacts the scores (performing even below the zero-shot baseline in some metrics). These findings indicate that the models fail to generalize when the test set contains samples following templates not provided in the training data. This exposes some limitations of recent LMs, i.e., even though LMs were fine-tuned for fill-in-the-mask DM prediction, they struggle to extrapolate to new templates requiring similar skills.

3.5. Error analysis

We focus our error analysis on the discourse-level senses: "arg marker" and "disc rel". The goal is to understand if the predicted DMs are aligned (regarding positioning) with the gold DMs.

Regarding zero-shot models, we make the following observations. BERT and XLM-R often predict DMs found in the DMs list for both senses, meaning they can predict conventional and sound DMs. However, there is some confusion between the labels "backward" and "forward" vs. "rebuttal", and "Comparison" vs. "Expansion", indicating that the models are not robust to challenging "edge cases" in the TADM dataset (i.e., different text sequences in the dataset where subtle changes in the content entail different argument structures and/or sets of DMs). BART-based models predict more diverse DMs, increasing the number of predictions not found in the DM lists. We observe less confusion between these labels, indicating that these models are more robust to the edge cases.

As the automatic evaluation metrics indicate, fine-tuned BERT ("all") performs better than the zero-shot models. Nonetheless, we still observe some confusion between the labels "backward" and "for-

ward" vs. "rebuttal" and "Contingency" vs. "Comparison", even though these confusions are much less frequent than in zero-shot models.

Some examples from the test set of the TADM dataset are shown in Table 8 (Appendix G). We also show the predictions made by the zero-shot models and fine-tuned `BERT` ("all") model in Table 9 (Appendix G).

3.6. Human evaluation

We conduct a human evaluation experiment to assess the quality of the predictions provided by the models in a zero-shot setting. Furthermore, we also aim to determine if the automatic evaluation metrics correlate with human assessments.

We selected 20 samples from the TADM dataset, covering different templates. For each sample, we get the predictions provided by each model analyzed in a zero-shot setting (i.e., `BERT`, `XLM-R`, `BART V1`, `BART V2`), resulting in a total of 80 samples; after removing repeated instances with the same prediction, each annotator analyzed 67 samples. Three annotators performed this assessment. Each annotator rates the prediction made by the model based on two criteria:

- Grammaticality: *Is the predicted content grammatically correct given the context?* The annotator is asked to rate with one of the following options: (-1): no, incorrect or out-of-context; (0): neutral or ambiguous; $(+1)$: yes, it is appropriate.
- Coherence: *Is the connotation of the predicted content correct/appropriate taking into account the surrounding context?* Options: (-1): incorrect, predicted content is in the opposite sense; (0): neutral or ambiguous; $(+1)$: correct, right sense.

Table 10 (Appendix H) shows some of the samples presented to the annotators and corresponding ratings.

We report inter-annotator agreement (IAA) using Cohen's κ metric [14], based on the scikit-learn [49] implementation. For "Grammaticality", we obtain a Cohen's κ score of 0.7543 (which corresponds to "substantial" agreement according to the widely used scale of values proposed by Landis and Koch [32]); for "Coherence", a "moderate" agreement of 0.5848. Overall, IAA scores indicate that human annotators can perform this task with reasonable agreement; however, assessing the "Coherence" criteria consistently (especially considering all the subtle changes in content that lead to different argument structures) is a challenging task that requires more cognitive effort. Analyzing disagreements, we observe that most of them occur when one of the annotators provides the rating "neutral or ambiguous" (0) while the other annotator considers either $(+1)$ or (-1). Additional details can be found in Appendix H, namely in Tables 11 and 12. We also perform a correlation analysis between criteria "Grammaticality" and "Coherence" to determine if the ratings provided by each annotator for these criteria are correlated. We obtain Pearson correlation coefficients of 0.0573, 0.0845, and 0.1945 (very low correlation). Therefore, we conclude that annotators can use the criteria independently.

To determine whether automatic evaluation metrics (Section 3.2) are aligned with human assessments, we perform a correlation analysis. To obtain a gold standard rating for each text sequence used in the human evaluation, we average the ratings provided by the three annotators. The results for the correlation analysis are presented in Table 2 (using the Pearson correlation coefficient). For the "Grammaticality" criterion, "retrofit embs" is the metric most correlated with human assessments, followed closely by "word embs". Regarding the "Coherence" criterion, "disc rel" is the most correlated metric. Intuitively, metrics based on discourse-level senses are more correlated with the "Coherence" criterion because they capture the discourse-level role that these words have in the sentence. On the other hand, these metrics

Table 2
Correlation analysis between human assessment criteria (columns) and automatic evaluation metrics (rows). Metric: Pearson correlation coefficient

metric	Grammaticality	Coherence
word embs	0.4757	0.2795
retrofit embs	**0.4907**	0.3243
sbert embs	0.2888	0.3519
arg marker	0.2075	0.3757
disc rel	−0.0406	**0.5421**

are less correlated with the "Grammaticality" criterion. We attribute this to the fact that a DM prediction can be correct in terms of "Grammaticality" but incorrect in terms of "Coherence" (e.g., examples where "Gram." = (+1) and "Coh." = (−1) in Table 10). In this situation, metrics based on discourse-level senses will indicate the DM prediction is incorrect, which is not appropriate through the lens of the "Grammaticality" criterion. This observation indicates that the proposed metrics are complementary: metrics based on discourse-level senses should be employed to evaluate the "Coherence" criterion, but embeddings-based metrics are more suitable for the "Grammaticality" criterion.

4. Data

In Section 3, based on experiments in a challenging synthetic testbed proposed in this work (the TADM dataset), we have shown that DMs play a crucial role in argument perception and that recent LMs struggle with DM prediction. We now move to more real-world experiments. The goal is to automatically augment a text with DMs (as performed in Section 5 in an end-to-end approach instead of the simplified "fill-in-the-mask" setup performed in Section 3) and assess the impact of the proposed approach in a downstream task (Section 6). In this section, we introduce the corpora used in the end-to-end DM augmentation (Section 5) and downstream task (Section 6) experiments. In Section 4.1, we describe three corpora containing annotations of argumentative content that are used in our experiments only for evaluation purposes. Then, in Section 4.2, we describe the corpora containing annotations of DMs. These corpora are used in our experiments to train Seq2Seq models to perform end-to-end DM augmentation.

4.1. Argument mining data

Persuasive Essays corpus (PEC). Contains token-level annotations of ADUs and their relations from student essays [63]. An argument consists of a claim and one (or more) premise(s), connected via argumentative relations: "Support" or "Attack". Arguments are constrained to the paragraph boundaries (i.e., the corresponding ADUs must occur in the same paragraph). Each ADU is labeled with one of the following: "Premise" (P), "Claim" (C), or "Major Claim" (MC). A paragraph might contain zero or more arguments. The distribution of token-level labels is: C (15%), P (45%), MC (8%), O (32%). PEC contains 402 essays, 1833 paragraphs, 1130 arguments, 6089 ADUs, and an average of 366 tokens per essay. PEC is a well-established and one of the most widely explored corpora for ArgMining tasks.

Microtext corpus (MTX). Contains token-level annotations of ADUs from short texts (six or fewer sentences) written in response to trigger questions, such as "Should one do X" [50]. Each microtext consists of one claim and several premises. Each ADU is labeled with one of the following: P or C. Note that all tokens are associated with an ADU (MTX does not contain O tokens). It contains 112 microtexts. The distribution of token-level labels is: C (18%) and P (82%).

Hotel reviews corpus (Hotel). Contains token-level annotations of ADUs from Tripadvisor hotel reviews [23,37]. Each sub-sentence in the review is considered a clause. Annotators were asked to annotate each clause with one of the following labels: "Background" (B), C, "Implicit Premise" (IP), MC, P, "Recommendation" (R), or "Non-argumentative" (O). It contains 194 reviews, with an average of 185 tokens per review. The distribution of token-level labels is: B (7%), C (39%), IP (8%), MC (7%), P (22%), R (5%), O (12%). We expect this to be a challenging corpus for ADU boundary detection and classification because: (a) it contains user-generated text with several abbreviations and grammatical errors; (b) the label space is larger; and (c) the original text is mostly deprived of explicit DMs.

4.2. Data with DM annotations

Template-based Arguments for Discourse Marker prediction (TADM). Dataset proposed in this paper to assess the capabilities of LMs to predict DMs. Each sample contains a claim and one (or two) premise(s). Every ADU is preceded by one DM that signals its role (further details are provided in Section 3.1).

Discovery. Provides a collection of adjacent sentence pairs $\langle s1, s2 \rangle$ and corresponding DM y that occurs at the beginning of $s2$ [61]. This corpus was designed for the Discourse Connective Prediction (DCP) task, where the goal is to determine y (from a fixed set of possible DMs) based on $s1$ and $s2$; e.g., $s1$ = "*The analysis results suggest that the HCI can identify incipient fan bearing failures and describe the bearing degradation process.*", $s2$ = "*The work presented in this paper provides a promising method for fan bearing health evaluation and prognosis.*", and y = "*overall,*". Input sequences are extracted from the Depcc web corpus [48], which consists of English texts collected from commoncrawl web data. This corpus differs from related work by the diversity of the DMs collected (a total of 174 different classes of DMs were collected, while related work collected 15 or fewer classes of DMs, e.g., the DisSent corpus [44]).

PDTB-3. Contains annotations of discourse relations for articles extracted from the Wall Street Journal (WSJ) [70].[6] These discourse relations describe the relationship between two discourse units (e.g., propositions or sentences) and are grounded on explicit DMs occurring in the text (explicit discourse relation) or in the adjacency of the discourse units (implicit discourse relation). For explicit relations, annotators were asked to annotate: the connective, the two text spans that hold the relation, and the sense it conveys based on the PDTB senses hierarchy. For implicit relations, annotators were asked to provide an explicit connective that best expresses the sense of the relation. This resource has been widely used for research related to discourse parsing.

5. End-to-end discourse marker augmentation

As detailed in Section 1, our goal is to automatically augment a text with DMs such that downstream task models can take advantage of explicit signals automatically added. To this end, we need to conceive models that can automatically (a) identify where DMs should be added and (b) determine which DMs should be added based on the context. We frame this task as a Seq2Seq problem. As input, the models receive a (grammatically sound) text sequence that might (or not) contain DMs. The output is the text sequence populated with DMs. For instance, for ArgMining tasks, we expect that the DMs should be added preceding ADUs.

[6] https://catalog.ldc.upenn.edu/LDC2019T05

Model. We employ recent Seq2Seq language models, namely: BART ("facebook/bart-base") [35] and T5 ("t5-base" and "t5-large") [54]. We use default "AutoModelForSeq2SeqLM" and "Seq2Seq Training Arguments" parameters provided by the HuggingFace library, except for the following: scheduler = "constant" and max training epochs = 5.

5.1. Evaluation

As previously indicated, explicit DMs are known to be strong indicators of argumentative content, but whether to include explicit DMs inside argument boundaries is an annotation guideline detail that differs across ArgMining corpora. Furthermore, in the original ArgMining corpora (e.g., the corpora explored in this work, Section 4.1), DMs are not directly annotated. We identify the gold DMs based on a heuristic approach proposed by Kuribayashi et al. [30]. For PEC, we consider as gold DM the span of text that precedes the corresponding ADU; more concretely, the span to the left of the ADU until the beginning of the sentence or the end of a preceding ADU is reached. For MTX, DMs are included inside ADU boundaries; in this case, if an ADU begins with a span of text specified in a DM list,[7] we consider that span of text as the gold DM and the following tokens as the ADU. For Hotel, not explored by Kuribayashi et al. [30], we proceed similarly to MTX. We would like to emphasize that following this heuristic-based approach to decouple DMs from ADUs (in the case of MTX and Hotel datasets) keeps sound and valid the assumption that DMs often precede the ADUs; this is already considered and studied in prior work [2,30], only requiring some additional pre-processing steps to be performed in this stage to normalize the ArgMining corpora in this axis.

To detokenize the token sequences provided by the ArgMining corpora, we use sacremoses.[8] As Seq2Seq models output a text sequence, we need to determine the DMs that were augmented in the text based on the corresponding output sequence. Similar to the approach detailed in Section 3.3, we use a diff-tool implementation, but in this case, we might have multiple mask tokens (one for each ADU). Based on this procedure, we obtain the list of DMs predicted by the model, which we can compare to the list of gold DMs extracted from the original input sequence.

To illustrate this procedure, consider the example in Fig. 1. The gold DMs for this input sequence are ["", "However", "", "In my opinion"]. An empty string means that we have an implicit DM (i.e., no DM preceding the ADU in the original input sequence); for the remaining, an explicit DM was identified in the original input sequence. The predicted DMs are ["Indeed", "However", "Furthermore", "In fact"], as illustrated in the underlined text in boxes (1) and (2) from Fig. 1.

In terms of evaluation protocol, we follow two procedures: (a) **Explicit DMs accuracy analysis**: based on the gold explicit DMs in the original input sequence, we determine whether the model predicted a DM in the same location and whether the prediction is correct (using the automatic evaluation metrics described in Section 3.2). For sense-based metrics, only gold DMs that can be mapped to some sense are evaluated. With this analysis, we aim to determine the quality of the predicted DMs (i.e., if they are aligned with gold DMs at the lexical and semantic-level). (b) **Coverage analysis**: based on all candidate DMs (explicit and implicit) that could be added to the input sequence (all elements in the gold DM list), we determine the percentage of DMs that are predicted. The aim of this analysis is to determine

[7] The longest occurring one, using the same list proposed by Kuribayashi et al. [30], which is composed of DMs that can be found in PEC and PDTB.

[8] https://github.com/alvations/sacremoses

Table 3

End-to-end DM augmentation results – explicit DMs accuracy analysis. Evaluated on the test set of PEC. Columns: "seq2seq model" indicates the pre-trained Seq2Seq LM, "fine-tune data" indicates the data used to fine-tune the Seq2Seq LM (details in Section 5.2), and the remaining columns correspond to the automatic evaluation metrics (Section 3.2). "Input data: removed DMs": version of the input data where all explicit DMs were removed, "Input data: original": original input data

seq2seq model	fine-tune data	word embs	retrofit embs	sbert embs	arg marker	disc rel
Input data: removed DMs						
BART-base	none	0.0075	0.0029	0.0021	0	0
	Discovery	0.4972	0.1969	0.2416	0.3572	0.2231
	TADM	**0.5878**	**0.2619**	0.2775	0.2475	0.2596
	PDTB	0.3805	0.2035	0.1871	0.2070	0.3236
BART-base	Discovery +	0.5038	0.2101	0.2563	0.3434	0.2150
T5-base	TADM + PDTB	0.5087	0.2372	**0.2817**	0.4290	**0.3393**
T5-large		0.4992	0.2308	0.2768	**0.4405**	0.3204
Input data: original						
BART-base	Discovery +	0.9075	0.8871	0.7998	0.1992	0.2383
T5-base	TADM + PDTB	**0.9125**	**0.8908**	0.8236	0.4835	0.5127
T5-large		0.8984	0.8781	**0.8275**	**0.5745**	**0.6247**

to which extent the model is augmenting the data with DMs in the correct locations (including implicit DMs, which could not be evaluated in (a)).[9]

For an input sequence, there may be multiple DMs to add; our metrics average over all occurrences of DMs. Then, we average over all input sequences to obtain the final scores, as reported in Table 3 (for explicit DMs accuracy analysis) and Table 4 (for coverage analysis).

Importantly, further changes to the input sequence might be performed by the Seq2Seq model (i.e., besides adding the DMs, the model might also commit/fix grammatical errors, capitalization, etc.), but we ignore these differences for the end-to-end DM augmentation assessment.[10]

5.2. Data preparation

In the following, we describe the data preparation for each corpus (we use the corpora mentioned in Section 4.2) to obtain the input and output sequences (gold data) for training the Seq2Seq models in the end-to-end DM augmentation task.

[9]Some notes regarding the nomenclature employed to name these two evaluation axes. For "Coverage analysis", we use the term "coverage" but "recall" could also be a viable option. Given that this assessment can be framed as a binary classification problem (i.e., whether a DM was predicted in an expected location), these terms can be used interchangeably. For "Explicit DMs accuracy analysis", we opted for the term "accuracy" instead of "precision". Given that the automatic evaluation metrics are very different in terms of what and how they measure different phenomena (i.e., similarity-based metrics and multi-class classification metrics for "arg marker" and "disc rel"), we believe that "accuracy" is a more general term and, consequently, more appropriate to summarize this evaluation axis. For the similarity-based metrics: "accuracy" refers to how close a measurement is to the true or accepted value. For multi-class classification, "accuracy" refers to the fraction of correct classifications out of all classifications.

[10]When we provide as input the original text (i.e., "Input data: original"), for the DM augmentation model `BART-base` fine-tuned on the combination of the corpora "Discovery + TADM + PDTB", we manually analyzed a total of 35 samples from the ArgMining datasets (PEC, MTX, and Hotel) where the model made at least one edit in an additional location (i.e., besides adding DMs preceding the ADUs). In all these cases, we observed minor edits being performed to the content of the ADUs, mostly related to specific grammatical edits that do not change the overall meaning (similar to the conclusions drawn in Appendix E for the fill-in-the-mask discourse marker prediction task).

Table 4

End-to-end DM augmentation results – coverage analysis. Columns: "seq2seq model" indicates the pre-trained Seq2Seq LM, "fine-tune data" indicates the data used to fine-tune the Seq2Seq LM (details in Section 5.2), and the remaining columns indicates the ArgMining corpus (evaluation is performed in the corresponding test sets). "Input data: removed DMs": version of the input data where all explicit DMs were removed, "Input data: original": original input data. Scores correspond to the percentage of DMs that are predicted (as detailed in Section 5.1)

seq2seq model	fine-tune data	PEC	MTX	Hotel
Input data: removed DMs				
BART-base	none	2.00	0	5.71
	Discovery	88.01	80.36	32.60
	TADM	88.61	**80.51**	40.21
	PDTB	61.61	54.93	22.69
BART-base	Discovery +	**88.72**	79.42	**63.28**
T5-base	TADM + PDTB	86.92	80.07	41.91
T5-large		85.84	73.99	28.53
Input data: original				
BART-base	Discovery +	**97.93**	**94.49**	**69.40**
T5-base	TADM + PDTB	95.61	92.25	45.71
T5-large		95.24	88.55	36.44

TADM dataset. For each sample, we generate a text sequence without any DM for the input sequence and another text sequence with DMs preceding all ADUs (e.g., text sequences 1 and 2 in Fig. 2) for the output sequence. To illustrate this process, the input sequence that corresponds to the output sequence 1 in Fig. 2 is "*Ecological concerns add further strain on the economy, we should introduce carbon taxes. Humanity must embrace clean energy in order to fight climate change*".

Discovery. For each sample, we employ the concatenation of the sentences without any DM in between for the input sequence (i.e., "s1. s2") and with the corresponding DM at the beginning of $s2$ for the output sequence (i.e., "s1. y s2").

PDTB-3. As input, we provide a version of the original text where all explicit DMs are removed. We also perform the following operations to obtain a grammatically sound text: (a) if the DM is not at the beginning of a sentence and if it is not preceded by any punctuation mark, we replace the DM with a comma – other options would be possible in specific cases, but we found the comma substitution to be a reasonable option (e.g., "*[. . .] this is a pleasant rally* ***but*** *it's very selective [. . .]*" is converted to "*[. . .] this is a pleasant rally, it's very selective [. . .]*"); (b) if the DM occurs at the beginning of a sentence, we uppercase the content that follows immediately after the removed DM. As output, we provide a version of the original text where the implicit DMs are also added. Adding implicit DMs also requires an extra pre-processing step, namely: if the DM occurs at the beginning of a sentence, we lowercase the content that follows the added DM.

5.3. Results

Setup. For evaluation purposes, we analyze the capability of Seq2Seq models to augment a given text with DMs using two versions of the input data: (a) the original input sequence ("Input data: original"), which contains the explicit DMs originally included in the text; (b) the input sequence obtained after the removal of all explicit DMs ("Input data: removed DMs"). The first setting can be seen as the standard application scenario of our proposed approach, where we ask the Seq2Seq model to augment a given

text, which might or might not contain explicit DMs. We expect the model to take advantage of explicit DMs to better capture the meaning of the text and to automatically augment the text with implicit DMs. The second setting is challenging because the Seq2Seq model is asked to augment a text deprived of explicit signals. To remove the explicit DMs from the original input sequence ("Input data: removed DMs"), we use the annotations of ADUs provided in ArgMining corpora. As described in Section 5.1, we follow a heuristic approach [30] to identify the gold DMs that precede each ADU. Then, we remove the corresponding DMs from the input sequence and perform the same operations described in Section 5.2 for PDTB-3 to obtain a grammatically sound text sequence (e.g., "*A great number of plants and animals died out* ***because*** *they were unable to fit into the new environment.*" is converted to "*A great number of plants and animals died out, they were unable to fit into the new environment*.").

Explicit DMs accuracy analysis. Table 3 details the results. The evaluation is performed on the test set of PEC, employing the automatic evaluation metrics described in Section 3.2. We do not employ the remaining corpora from Section 4.1 because the percentage of explicit DMs is relatively low.

We start our analysis with the "Input data: removed DMs" setting. First, we observe that the pre-trained `BART-base` model underperforms in the DM augmentation task in a zero-shot setting ("none" in the column "fine-tune data") because it will not automatically augment the text with DMs without being explicitly trained to do it. Then, we compare the scores obtained when fine-tuning the pre-trained `BART-base` model on each corpus individually ("Discovery", "TADM", and "PDTB"). We observe that the best scores on the: (a) embeddings-based metrics (i.e., "word embs", "retrofit embs", and "sbert embs") are obtained when fine-tuning `BART-base` on TADM, which we attribute to the restricted set of DMs used in the training data and, consequently, the predictions made by the model are more controlled towards well-known DMs; (b) "disc rel" metric is obtained fine-tuning on PDTB, which indicates that this corpus is relevant to improve the models on the "Coherence" axis; (c) "arg marker" metric is obtained fine-tuning on Discovery. We also provide results when fine-tuning on the combination of the three corpora in the training set; we consider `BART-base`, `T5-base`, and `T5-large` pre-trained models. We make the following observations: (i) surprisingly, `BART-base` performs worse when fine-tuned on the combination of all datasets compared to the best individual results; (ii) `T5-base` is superior to `BART-base` in all metrics; (iii) `T5-base` performs better than `T5-large` in most metrics (except for "arg marker"), indicating that larger models do not necessarily perform better in this task.

Regarding the "Input data: original" setting, we analyze the results obtained after fine-tuning on the combination of the three corpora. As expected, we obtain higher scores across all metrics compared to "Input data: removed DMs", as the Seq2Seq model can explore explicit DMs (given that we frame it as a Seq2Seq problem, the model might keep, edit, or remove explicit DMs) to capture the semantics of text and use this information to improve on implicit DM augmentation. `BART-base` underperforms in this setting compared to the T5 models. We observe higher variations in the scores for the metrics "arg marker" and "disc rel", with `T5-large` obtaining remarkable improvements, almost 10 percentage points above `T5-base`, which itself is 30 points above `BART-base`.

Coverage analysis. Detailed results are provided in Table 4. The evaluation is performed on the test set of each ArgMining corpus described in Section 4.1. For reference, the results obtained in the original data (i.e., containing only the original explicit DMs, which corresponds to "Input data: original" without the intervention of any Seq2Seq model) are: 73% for PEC, 44% for MTX, and 15% for Hotel. We explore the same input data settings and Seq2Seq pre-trained models, and fine-tune with data previously detailed for the explicit DM accuracy analysis.

Analyzing the results obtained with the "Input data: removed DMs" setting, we observe that: `BART-base` employed in a zero-shot setting underperforms the task (because it will not augment the text with DMs); fine-tuning on individual corpora improves the scores (Hotel seems to be the most challenging corpus); a model trained solely on PDTB obtains the lowest scores across all corpora, while Discovery and TADM perform on par in PEC and MTX, but the model trained on TADM stands out with higher scores on Hotel. The scores obtained after fine-tuning on individual corpora are superior to the reference values reported for the original data (except for PDTB on PEC), meaning that the Seq2Seq models successfully increase the coverage of ADUs being preceded by explicit DMs (even departing from the input data deprived of DMs, i.e., "Input data: removed DMs" setting). Combining the corpora positively impacts the scores on Hotel (23 percentage points above best individual results), with similar scores obtained on PEC and MTX. Surprisingly, `T5-large` again obtains lower scores.

For the "Input data: original" setting, we again obtain higher scores. These improvements are smaller for Hotel because the original text is mostly deprived of explicit DMs. Finally, we observe that in this setting, we can obtain very high coverage scores across all corpora: 98% for PEC, 95% for MTX, and 69% for Hotel.

5.4. Comparison with ChatGPT

ChatGPT[11] [34] is a popular large language model built on top of the GPT-3.5 series [8] and optimized to interact in a conversational way. Even though ChatGPT is publicly available, interacting with it can only be done with limited access. Consequently, we are unable to conduct large-scale experiments and fine-tune the model on specific tasks. To compare the zero-shot performance of ChatGPT in our task, we run a small-scale experiment and compare the results obtained in the "Input data: removed DMs" setting with the models presented in Section 5.3. We sampled 11 essays from the test set of the PEC (totaling 60 paragraphs) for this small-scale experiment, following the same evaluation protocol described in Section 5.1. ChatGPT is trained to follow an instruction expressed in a prompt, targeting a detailed response that answers the prompt. To this end, we provide the prompt "add all relevant discourse markers to the following text: {}" for each paragraph. Then, we collect the answer provided by ChatGPT as the predicted output sequence. Detailed results can be found in Appendix I. Even though our fine-tuned models surpass ChatGPT in most of the metrics (except for "disc rel"), especially in terms of coverage, it is remarkable that ChatGPT, operating in a zero-shot setting, is competitive. With some fine-tuning, better prompt-tuning or in-context learning, we believe that ChatGPT (and similar LLMs) might excel in the proposed DM augmentation task.

6. Downstream task evaluation

We assess the impact of the end-to-end DM augmentation approach detailed in Section 5 on an ArgMining downstream task, namely: ADU identification and classification. We operate on the token level with the label set: $\{O\} \cup \{B, I\} \times T$, where T is a corpus-specific set of ADU labels. For instance, for PEC, $T = \{P, C, MC\}$. This subtask is one of the most fundamental in the ArgMining process and considered in many other ArgMining studies [18,41,59]. Its reduced complexity compared to tasks that also include relation identification makes a subsequent analysis of the impact of the proposed approach easier.

[11] https://openai.com/blog/chatgpt/

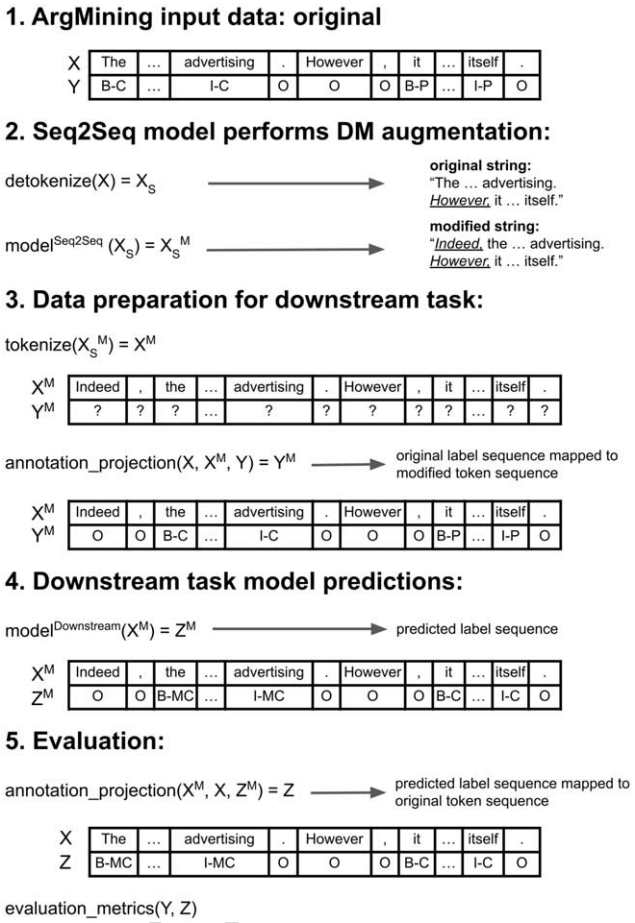


Fig. 3. Downstream task experimental setup process (details provided in Section 6.1).

6.1. Experimental setup

We assess the impact of the proposed DM augmentation approach in the downstream task when the Seq2Seq models (described in Section 5) are asked to augment a text based on two versions (similar to Section 5.3) of the input data: (a) the original input sequence ("Input data: original"); (b) the input sequence obtained after the removal of all explicit DMs ("Input data: removed DMs").

In Fig. 3, we illustrate the experimental setup process using "Input data: original" (step 1), where X corresponds to the original token sequence and Y to the original label sequence (as provided in the gold ArgMining annotations). The process is similar for "Input data: removed DMs".

In step 2 (Fig. 3), a Seq2Seq model performs DM augmentation. Since Seq2Seq models work with strings and the input data is provided as token sequences, we need to detokenize the original token sequence (resulting in X_S in Fig. 3). All tokenization and detokenization operations are performed using sacremoses. At the end of this step, we obtain the output provided by the Seq2Seq model (i.e., X_S^M, the text augmented with DMs), which will be used as the input data for the downstream task model (the model trained and evaluated on the downstream task) in the following steps.

Given that the output provided by the Seq2Seq model is different from the original token sequence X (based on which we have the gold ArgMining annotations), we need to map the original label sequence (Y) to the modified token sequence (i.e., X^M, the token sequence obtained after tokenization of the Seq2Seq output string X_S^M). To this end, in step 3 (Fig. 3), we employ an annotation projection procedure, detailed in Appendix J. Based on this annotation projection procedure, we train the downstream model using the modified token sequence (X^M) and the corresponding label sequence obtained via annotation projection (i.e., Y^M, the original label sequence mapped to the modified token sequence). Then, using the trained model, we obtain, in step 4 (Fig. 3), the predictions for the test set (i.e., Z^M, which also contains modified sequences).

For a fair comparison between different approaches, in step 5 (Fig. 3), we map back the predicted label sequence (Z^M) to the original token sequence (i.e., Z corresponds to the predicted label sequence mapped to the original token sequence), using the same annotation projection procedure. Consequently, despite all the changes made by the Seq2Seq model, we ensure that the downstream task evaluation is performed on the same grounds for each approach. This is crucial to obtain insightful token-level and component-level (i.e., ADUs in this downstream task) metrics. As evaluation metrics, we use the following: (a) seqeval[12] is a popular framework for sequence labeling evaluation typically used to evaluate the performance on chunking tasks such as named-entity recognition and semantic role labeling; (b) flat token-level macro-F1 as implemented in scikit-learn [49].

For the downstream task, we employ a BERT model ("bert-base-cased") following a typical sequence labeling approach. We use default "AutoModelForTokenClassification" and "TrainingArguments" parameters provided by the HuggingFace library, except for the following: learning rate $= 2e-5$, weight decay $= 0.01$, max training epochs $= 50$, and evaluation metric (to select the best epoch based on the dev set) is token-level macro f1-score (similar to prior work, e.g., Schulz et al. [59]).

6.2. Results

Table 5 shows the results obtained for the downstream task. The line identified with a "none" in the column "DM augmentation model" refers to the scores obtained by a baseline model in the corresponding input data setting, in which the downstream model receives as input the data without the intervention of any DM augmentation Seq2Seq model; for the remaining lines, the input sequence provided to the downstream model was previously augmented using a pre-trained Seq2Seq model (`BART-base`, `T5-base`, or `T5-large`) fine-tuned on the combination of the three corpora described in Section 4.2 ("Discovery + TADM + PDTB"). For "Input data: original", the scores in "none" correspond to the current state-of-the-art using a BERT model. For "Input data: removed DMs", the scores in "none" correspond to a very challenging setup for sequence labeling models because they are asked to perform the downstream task without explicit signals.

We start our analysis by comparing the baseline models, i.e., rows (1) and (A). When the DMs are removed from the data, row (1), we observe an expected drop in the scores in all the metrics on the PEC and MTX corpora: the task is more challenging without DMs. Given that the original data in the Hotel corpus is already scarce regarding DMs, we observe slightly lower scores for the metric "token macro f1" with the removal of DMs. Surprisingly, we observe higher scores (by a small margin) for the remaining metrics.

Most of the results obtained from DM augmentation models that receive as input the original data ("Input data: original"; rows (A, B, C, D)) are superior to the scores obtained in the setting "Input data:

[12] https://huggingface.co/spaces/evaluate-metric/seqeval

Table 5

Downstream task (ADU identification and classification) results. Columns: "DM augmentation model" indicates the pre-trained Seq2Seq model that was fine-tuned on "Discovery + TADM + PDTB" to augment the input data with DMs; for each ArgMining corpus (PEC, MTX, and Hotel), we report the results obtained by a BERT model fine-tuned on the downstream task using the metrics: "seqeval f1", flat token-level accuracy ("token acc"), and macro-F1 ("token macro f1"). "Input data: removed DMs": version of the input data where all explicit DMs were removed, "Input data: original": original input data

DM augmentation model		PEC			MTX			Hotel		
		seqeval f1	token acc	token macro f1	seqeval f1	token acc	token macro f1	seqeval f1	token acc	token macro f1
Input data: removed DMs										
(1)	none	0.6241 (± 0.007)	**0.8294** (± 0.005)	0.7457 (± 0.003)	0.6922 (± 0.029)	0.8993 (± 0.005)	0.8409 (± 0.011)	**0.3341** (± 0.013)	0.5016 (± 0.018)	**0.4855** (± 0.003)
(2)	BART-base	0.6301 (± 0.009)	0.8252 (± 0.004)	0.7435 (± 0.007)	0.7114 (± 0.025)	0.8995 (± 0.015)	0.8437 (± 0.018)	0.2703 (± 0.008)	0.4907 (± 0.004)	0.4637 (± 0.009)
(3)	T5-base	**0.6397** (± 0.007)	0.8247 (± 0.003)	0.7438 (± 0.002)	**0.7491** (± 0.033)	0.9129 (± 0.007)	0.8630 (± 0.008)	0.3121 (± 0.022)	**0.5540** (± 0.028)	0.4764 (± 0.019)
(4)	T5-large	0.6284 (± 0.020)	0.8258 (± 0.011)	**0.7459** (± 0.010)	0.6988 (± 0.067)	**0.9140** (± 0.017)	**0.8671** (± 0.021)	0.2902 (± 0.020)	0.5038 (± 0.019)	0.4307 (± 0.013)
Input data: original										
(A)	none	**0.6785** (± 0.013)	**0.8557** (± 0.003)	**0.7871** (± 0.004)	0.7489 (± 0.037)	**0.9193** (± 0.009)	0.8674 (± 0.019)	0.3120 (± 0.010)	0.4928 (± 0.009)	**0.4948** (± 0.023)
(B)	BART-base	0.6569 (± 0.022)	0.8466 (± 0.006)	0.7758 (± 0.004)	0.7398 (± 0.021)	0.9150 (± 0.004)	**0.8676** (± 0.009)	**0.3345** (± 0.008)	**0.5021** (± 0.008)	0.4762 (± 0.013)
(C)	T5-base	0.6486 (± 0.014)	0.8396 (± 0.010)	0.7699 (± 0.010)	0.7550 (± 0.014)	0.9136 (± 0.012)	0.8554 (± 0.004)	0.3183 (± 0.043)	0.4841 (± 0.004)	0.4666 (± 0.023)
(D)	T5-large	0.6264 (± 0.020)	0.8294 (± 0.009)	0.7556 (± 0.007)	**0.7719** (± 0.021)	0.9157 (± 0.001)	0.8665 (± 0.006)	0.2885 (± 0.008)	0.4645 (± 0.027)	0.4443 (± 0.011)

removed DMs" (rows (1, 2, 3, 4)). However, even though in Section 5 we observed clear improvements in all metrics for the setting "Input data: original", these improvements are reflected with limited effect in the downstream task.

Comparing row (1), which is the baseline with DMs removed, to rows (2, 3, 4), which give the results after adding DMs with the DM augmentation models previously described, we observe: (i) consistent improvements for the MTX dataset, i.e., the results of (2, 3, 4) are better than (1) in all cases; (ii) for PEC, all rows (1, 2, 3, 4) have mostly similar scores across the metrics; (iii) only `T5-base` clearly improves, and only for "token accuracy", over (1) for Hotel.

Comparing row (*A*), which is the baseline with original data, to (*B*, *C*, *D*), which give the results after performing DM augmentation on top of the original data with the Seq2Seq models previously described, we observe: (i) the most consistent improvements are obtained again for the MTX dataset, where we observe a 3 percentage point improvement in "seqeval f1" for `T5-large` over the baseline; (ii) `BART-base` improves upon the baseline for Hotel according to 2 out of 3 metrics; (iii) there are no improvements for PEC, the baseline performs better according to all three metrics (we attribute this to the high percentage of explicit DMs already included in the original data on PEC, i.e., 73% of the ADUs are preceded by an explicit DM, as indicated in Section 5.3).

To summarize, we observe that in a downstream task, augmenting DMs automatically with recent LMs can be beneficial in some, but not all, cases. We believe that with the advent of larger LMs (such as the recent ChatGPT), the capability of these models to perform DM augmentation can be decisively improved

in the near future, with a potential impact on downstream tasks (such as ArgMining tasks). Overall, our findings not only show the impact of a DM augmentation approach for a downstream ArgMining task but also demonstrate how the proposed approach can be employed to improve the readability and transparency of argument exposition (i.e., introducing explicit signals that clearly unveil the presence and positioning of the ADUs conveyed in the text).

Finally, we would like to reinforce that the DM augmentation models were fine-tuned on datasets (i.e., "Discovery + TADM + PDTB") that (a) were not manually annotated for ArgMining tasks and (b) are different from the downstream task evaluation datasets (i.e., PEC, MTX, and Hotel). Consequently, our results indicate that DM augmentation models can be trained on external data to automatically add useful DMs (in some cases, as previously detailed) for downstream task models, despite the differences in the DM augmentation training data and the downstream evaluation data (e.g., domain shift).

6.3. Error analysis

We manually sampled some data instances from each ArgMining corpora (Section 4.1) and analyzed the predictions (token-level, for the ADU identification and classification task) made by the downstream task models. Furthermore, we also analyzed the DMs automatically added by the Seq2Seq models, assessing whether it is possible to find associations between the DMs that precede the ADUs and the corresponding ADU labels.

In Appendix K, we provide a detailed analysis for each corpus and show some examples. Overall, we observe that DM augmentation models performed well in terms of coverage, augmenting the text with DMs at appropriate locations (i.e., preceding the DMs). This observation is in line with the conclusions taken from the "Coverage analysis" in Section 5.3. However, we observed that the presence of some DMs that are commonly associated as indicators of specific ADU labels (e.g., "because" and "moreover" typically associated to P) are not consistently used by the downstream model to predict the corresponding ADU label accordingly (i.e., the predicted ADU label varies in the presence of these DMs). We attribute this to the lack of consistency (we observed, for all corpora, that some DMs are associated to different ADU labels) and variability (e.g., on PEC, in the presence of augmented DMs, the label MC does not contain clear indicators; in the original text, these indicators are available and explored by the downstream model) of augmented DMs. We conclude that these limitations in the quality of the predictions provided by the DM augmentation models conditioned the association of DMs and ADU labels that we expected to be performed by the downstream model.

Based on this analysis, our assessment is that erroneous predictions of DMs might interfere with the interpretation of the arguments exposed in the text (and, in some cases, might even mislead the downstream model). This is an expected drawback from a pipeline architecture (i.e., end-to-end DM augmentation followed by the downstream task). However, on the other hand, the potential of DM augmentation approaches is evident, as the presence of coherent and grammatically correct DMs can clearly improve the readability of the text and of argument exposition in particular (as illustrated in the detailed analysis provided in Appendix K).

7. Related work

Argument mining. Given the complexity of the task, it is common to divide the ArgMining task in a set of subtasks [50], namely: ADU identification, ADU classification (e.g., premise vs. claim), Argumentative Relation Identification (ARI, e.g., link vs. no-link), and Argumentative Relation Classification

(ARC, e.g., support vs. attack). In this paper, we focus on ADU identification and classification as downstream tasks (Section 6). The standard BiLSTM with a CRF output layer emerged as the state-of-the-art architecture for token-level sequence tagging, including argument mining [9,17,59]. Current state-of-the-art on ADU identification and classification employs BERT [15] or Longformer [3] as base encoders (in some cases, with a CRF layer on top), typically accompanied with specific architectures to tackle a target corpus or task-specific challenges [16,24,41,69,71]. We follow these recent trends by employing a BERT-based sequence labeling model. Since our goal is to assess the impact of the proposed DM augmentation approach, we keep the architecture as simple and generic as possible (standard BERT encoder with a token classification head), but competitive with recent state-of-the-art (as detailed in Section 6).

Some prior work also studies ArgMining across different corpora. Given the variability of annotation schemas, dealing with different conceptualizations (such as tree vs. graph-based structures, ADU and relation labels, ADU boundaries, among others) is a common challenge [2,10,22]. Besides the variability of annotated resources, ArgMining corpora tend to be small [41]. To overcome these challenges, some approaches explored transfer learning: (a) across different ArgMining corpora [41,52,59]; (b) from auxiliary tasks, such as discourse parsing [1] and fine-tuning pre-trained LMs on large amounts of unlabeled discussion threads from Reddit [16]; and (c) from ArgMining corpora in different languages [18,19,57,66]. Exploring additional training data is pointed out as beneficial across different subtasks, especially under low-resource settings; however, domain-shift and differences in annotation schemas are typically referred to as the main challenges. Our approach differs by proposing DM augmentation to improve the ability of ArgMining models across different genres, without requiring to devise transfer learning approaches to deal with different annotation schemas: given that the DM augmentation follows a text-to-text approach, we can employ corpus-specific models to address the task for each corpus.

The role of discourse context. As a discourse parsing task, prior work on ArgMining looked at the intersection between argumentation structures and existing discourse parsing theories (e.g., RST, PDTB), with several studies pointing out that improvements can be obtained for ArgMining tasks by incorporating insights from related discourse parsing tasks [25,27]. From the state-of-the-art in discourse parsing tasks, it is well known that discourse markers play an important role as strong indicators for discourse relations [7,39]. In the field of ArgMining, such lexical clues have also been explored in prior work, either via handcrafted features [46,62,63] or encoding these representations in neural-based architectures [2,30,56]. Including DM in their span representations, Bao et al. [2] report state-of-the-art results for ADU classification, ARI, and ARC. These works rely on the presence of explicit DMs anteceding ADUs, which is a viable assumption for some of the ArgMining corpora containing texts written in English. To obtain a system that is robust either in the presence or absence of such lexical clues, we propose to automatically augment the text with the missing DMs using state-of-the-art Seq2Seq models. Our proposal complements prior work findings (e.g., including DMs in span representations improves performance across different subtasks) as we propose a text-to-text approach that can be employed to augment the input text provided to state-of-the-art ArgMining models.

Aligned with our proposal, Opitz [45] frames ARC as a plausibility ranking prediction task. The notion of plausibility comes from adding DMs (from a handcrafted set of 4 possible DM pairs) of different categories (support and attack) between two ADUs and determining which of them is more plausible. They report promising results for this subtask, demonstrating that explicitation of DMs can be a feasible approach to tackle some ArgMining subtasks. We aim to go one step further by: (a) employing language models to predict plausible DMs (instead of using a handcrafted set of DMs) and (b) proposing a more realistic DM augmentation scenario, where we receive as input raw text and we do not assume that the ADU boundaries are known.

However, relying on these DMs also has downsides. In a different line of work, Opitz and Frank [46] show that the models they employ to address the task of ARC tend to focus on DMs instead of the actual ADU content. They argue that such a system can be easily fooled in cross-document settings (i.e., ADUs belonging to a given argument can be retrieved from different documents), proposing a context-agnostic model that is constrained to encode only the actual ADU content as an alternative. We believe that our approach addresses these concerns as follows: (a) for the ArgMining tasks addressed in this work, arguments are constrained to document boundaries (cross-document settings are out of scope); (b) given that the DM augmentation models are automatically employed for each document, we hypothesize that the models will take into account the surrounding context and adapt the DMs predictions accordingly (consequently, the downstream model can rely on them).

Explicit vs. implicit relations in discourse parsing. In discourse parsing, it is well-known that there exists a clear gap between explicit (relations that are marked explicitly with a DM) and implicit (relation between two spans of text exists, but is not marked explicitly with a DM) relation classification, namely, 90% vs. 50% of accuracy (respectively) in 4-way classification (as indicated by Shi and Demberg [60]). To improve discourse relation parsing, several works focused on enhancing their systems for implicit relation classification: removing DMs from explicit relations for implicit relation classification data augmentation [6,58]; framing explicit vs. implicit relation classification as a domain adaptation problem [26,53]; learning sentence representations by exploring automatically collected large-scale datasets [44,61]; multi-task learning [31,43]; automatic explicitation of implicit DMs followed by explicit relation classification [28,29,60].

To close the gap between explicit and implicit DMs, our approach follows the line of work on explicitation. However, we work in a more challenging scenario, where the DM augmentation is performed at the paragraph level following an end-to-end approach (i.e., from raw text to a DM-augmented text). Consequently, our approach differs from prior work in multiple ways: (a) we aim to explicitate all discourse relations at the paragraph level, while prior work focuses on one relation at a time [28,29,60,64] (our models can explore a wider context window and take into account the interdependencies between different discourse relations), and (b) we do not require any additional information, such as prior knowledge regarding discourse units boundaries (e.g., clauses or sentences) [28,29,60,64] or the target discourse relations [64].

8. Conclusions

In this paper, we propose to automatically augment a text with DMs to improve the robustness of ArgMining systems across different genres.

First, we describe a synthetic template-based test suite created to assess the capabilities of recent LMs to predict DMs and whether LMs are sensitive to specific semantically-critical edits in the text. We show that LMs underperform this task in a zero-shot setting, but the performance can be improved after some fine-tuning.

Then, we assess whether LMs can be employed to automatically augment a text with coherent and grammatically correct DMs in an end-to-end setting. We collect a heterogeneous collection of DM-related datasets and show that fine-tuning LMs in this collection improves the ability of LMs in this task.

Finally, we evaluate the impact of augmented DMs performed by the proposed end-to-end DM augmentation models on the performance of a downstream model (across different ArgMining corpora). We

obtained mixed results across different corpora. Our analysis indicates that DM augmentation models performed well in terms of coverage; however, the lack of consistency and variability of the augmented DMs conditioned the association of DMs and ADU labels that we expected to be performed by the downstream model.

In future work, we would like to assess how recent LLMs perform in these tasks. Additionally, we would like to increase and improve the variability and quality of the heterogeneous collection of data instances used to fine-tune the end-to-end DM augmentation models (possibly including data related to ArgMining tasks that might inform the models about DMs that are more predominant in ArgMining domains), as improving in this axis might have a direct impact in the downstream task performance.

We believe that our findings are evidence of the potential of DM augmentation approaches. DM augmentation models can be deployed to improve the readability and transparency of arguments exposed in written text, such as embedding this approach in writing assistant tools.

9. Limitations

One of the anchors of this work is evidence from prior work that DMs can play an important role to identify and classify ADUs; prior work is mostly based on DMs preceding the ADUs. Consequently, we focus on DMs preceding the ADUs. We note that DMs following ADUs might also occur in natural language and might be indicative of ADU roles. However, this phenomenon is less frequent in natural language and also less studied in related work [18,30,61].

The TADM dataset proposed in Section 3.1 follows a template-based approach, instantiated with examples extracted from the CoPAs provided by Bilu et al. [4]. While some control over linguistic phenomena occurring in the dataset was important to investigate our hypothesis, the downside is a lack of diversity. Nonetheless, we believe that the dataset contains enough diversity for the purposes studied in this work (e.g., different topics, several parameters that result in different sentence structures, etc.). Future work might include extending the dataset with a wider spectrum of DMs, data instances, templates, and argument structures (e.g., including linked and serial argument structures besides the convergent structures already included).

Our proposed approach follows a pipeline architecture: end-to-end DM augmentation followed by the downstream task. Consequently, erroneous predictions made by the DM augmentation model might mislead the downstream task model. Furthermore, the end-to-end DM augmentation employs a Seq2Seq model. Even though these models were trained to add DMs without changing further content, it might happen in some cases that the original ADU content is changed by the model. We foresee that, in extreme cases, these edits might lead to a different argument content being expressed (e.g., changing the stance, adding/removing negation expressions, etc.); however, we note that we did not observe this in our experiments. In a few cases, we observed minor edits being performed to the content of the ADUs, mostly related to grammatical corrections.

We point out that despite the limited effectiveness of the proposed DM augmentation approach in improving the downstream task scores in some settings, our proposal is grounded on a well-motivated and promising research hypothesis, solid experimental setup, and detailed error analysis that we hope can guide future research. Similar to recent trends in the community (Insights NLP workshop [65], ICBINB Neurips workshop and initiative,[13] etc.), we believe that well-motivated and well-executed research can

[13] http://icbinb.cc/

also contribute to the progress of science, going beyond the current emphasis on state-of-the-art results.

Acknowledgements

Gil Rocha was supported by a PhD grant (SFRH/BD/140125/2018) from Fundação para a Ciência e a Tecnologia (FCT). This work was supported by Base Funding (UIDB/00027/2020) and Programmatic Funding (UIDP/00027/2020) of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through FCT/MCTES (PIDDAC). The NLLG group is supported by the BMBF grant "Metrics4NLG" and the DFG Heisenberg Grant EG 375/5–1. We would like to thank the anonymous reviewers for their valuable and thorough feedback that greatly helped improve the current version of the paper.

Appendix A. TADM dataset – templates

The templates are based on a set of configurable parameters, namely:

- number of ADUs $\in \{2, 3\}$: sample might contain 2 ADUs following the structure "dm_1 X_1, dm_2 X_2.", where one of the ADUs (X_1 or X_2) is a claim and the other a premise; or contain 3 ADUs (claim and both premises) following the structure "dm_1 X_1, dm_2 X_2. dm_3, X_3.";
- stance role $\in \{original, opposite\}$: each sample contains a single claim, which might employ the "original" or "opposite" stance;
- claim position $\in \{1, 2\}$: the claim is always in the first sentence,[14] either in the beginning (1) or end (2) of the sentence;
- premise role $\in \{support, attack\}$: only used when "number of ADUs" $= 2$; dictates which of the premises is chosen;
- supportive premise position $\in \{1, 2\}$: only used when "number of ADUs" $= 3$, indicates whether the supportive premise should occur before (1) the attacking premise or after (2);
- prediction type $\in \{dm_1, dm_2, dm_3\}$: let dm_i be the option chosen, then the mask token will be placed in DM preceding the ADU in position i (if "number of ADUs" $= 2$ only dm_1 and dm_2 are allowed).

Appendix B. TADM dataset – DMs set

DMs are added based on the role of the ADU that they precede, using the following fixed set of DMs. If preceding the claim, then we add one the following: "I think that", "in my opinion", or "I believe that". For the supportive premise, if in position dm_3 we add one of the following: "moreover", "furthermore", or "indeed". Otherwise, one of the following: "because", "since", or "given that". For the attacking premise, in dm_3 we add: "however", "on the other hand", or "conversely". Otherwise, "although", "even though", or "even if".

[14] Considering the set of DMs employed in the TADM dataset (Appendix B), an input sequence with two premises in a single sentence followed by an isolated claim in the second sentence (not allowed by the "claim position" parameter) results in an unusual input sequence. Expanding the set of DMs might relax this constraint and allow different templates to be included in the TADM dataset.

Appendix C. TADM dataset – instantiation procedure based on CoPAs

CoPAs are sets of propositions that are often used when debating a recurring theme (e.g., the premises mentioned in Section 3.1 and used in Fig. 2 are related to the theme "Clean energy"). For each CoPA, Bilu et al. [4] provide two propositions that people tend to agree as supporting different points of view for a given theme. We use these propositions as supportive and attacking premises towards a given claim. Each CoPA is also associated with a set of motions to which the corresponding theme is relevant. A motion is defined as a pair $\langle action, topic \rangle$, where an action is a term coming from a closed set of allowed actions (e.g., abolish, adopt, legalize, etc.), and a topic is a Wikipedia title. For example, for the theme "Clean energy", we can find the motion $\langle introduce, carbon\ taxes \rangle$, which can be written as "we should introduce carbon taxes". We use these motions as claims in our TADM dataset. Based on these instantiations and the set of templates, we can generate different samples that resemble real-world arguments. The "opposite stance" is not provided in the original resources from Bilu et al. [4]. For a specific motion, we manually selected the action (from the set of allowed actions) that could be employed as "opposite stance" (e.g., $\langle abolish, carbon\ taxes \rangle$).

Appendix D. Automatic evaluation metrics

The evaluation metrics employed to assess the quality of the predicted DMs are the following:

- word embeddings text similarity ("word embs"): Cosine similarity using an average of word vectors. Based on pre-trained embeddings "en_core_web_lg" from Spacy library;[15]
- retrofitted word embeddings text similarity ("retrofit embs"): Cosine similarity using an average of word vectors. Based on pre-trained embeddings from LEAR;[16]
- sentence embeddings text similarity ("sbert embs"): we use the pre-trained sentence embeddings "all-mpnet-base-v2" from SBERT library [55], indicated as the model with the highest average performance on encoding sentences over 14 diverse tasks from different domains. To compare gold and predicted DMs representations, we use cosine similarity;
- argument marker sense ("arg marker"): list of 115 DMs from Stab and Gurevych [63]. Senses are divided in the categories: "forward", "backward", "thesis", and "rebuttal". Each gold and predicted DM is mapped to one of the senses based on a strict lexical match with the list of DMs available for each sense. If the DM is not matched, then we assign the label "none". If the gold DM is "none", we do not consider this instance in the evaluation (the DM is out of the scope for the list of DMs available in the senses list, so we cannot make a concrete comparison with the predicted DM);
- discourse relation sense ("disc rel"): we use a lexicon of 149 English DMs called "DiMLex-Eng" [12]. These DMs were extracted from PDTB 2.0 [51], RST-SC [13], and Relational Indicator List [5]. Each DM maps to a set of possible PDTB senses. For a given DM, we choose the sense that the DM occurs more frequently. PDTB senses are organized hierarchically in 3 levels (e.g., the DM "consequently" is mapped to the sense "Contingency.Cause.Result"). In this work, we consider only the first level of the senses (i.e., "Comparison", "Contingency", "Expansion", and "Temporal") as a simplification and to avoid the propagation of errors between levels (i.e., an error in level 1 entails an error in level 2, and so on). Each prediction is mapped to one of the senses based on a strict lexical match. If the word is not matched, then we assign the label "none";

[15] https://spacy.io/
[16] https://github.com/nmrksic/lear

Appendix E. Fill-in-the-mask discourse marker prediction – BART V2 additional edits analysis

As detailed in Section 3.3, for the fill-in-the-mask discourse marker prediction task, we report the results of BART following a Seq2Seq setting (dubbed as `BART V2`). In this setting, the output sequence might contain further differences as compared to the input sequence (i.e., the model might perform further edits to the sequence besides mask-filling). For the fill-in-the-mask discourse marker prediction task reported in Section 3.4, 74 samples from the total of 600 samples in the test set of the TADM dataset (i.e., 12.3% of the samples) contain further edits besides mask-filling. We manually analyzed 40 of these samples, concluding that in all these cases the additional edits are minor and mostly related to specific grammatical edits that do not change the overall meaning of the content. Table 6 illustrates some samples where the `BART V2` model performed further edits to the text sequence besides mask-filling.

Table 6

Fill-in-the-mask discourse marker prediction – some samples where zero-shot Seq2Seq BART V2 model performed further edits to the text sequence besides mask-filling. "Masked sentence" corresponds to the sample extracted from the TADM dataset (in parentheses, we show the gold <mask> tokens), "Output sequence" to the output obtained from BART V2 (in parentheses, we show the predicted <mask> tokens obtained from "Output sequence", as detailed in Section 3.3). The underlined text highlights differences between "Masked sentence" and "Output sequence"

Masked sentence:
I think that big governments should be abandoned, *<mask>* public utility is best served by actions coordinated by central government. Moreover, public interest is *best* served and propelled by voluntary interactions, and not ones dictated by government.
(gold <mask> = "although")
Output sequence:
I think that big governments should be abandoned, *because* public utility is best served by actions coordinated by central government. Moreover, public interest is *better* served and propelled by voluntary interactions, and not ones dictated by government.
(predicted <mask> = "because")
Masked sentence:
<mask> some rights and freedoms need to be limited in the interest of national security, I think that we should end the use of mass surveillance. Moreover, security does not justify brushing aside fundamental right*s* and freedoms.
(gold <mask> = "Although")
Output sequence:
While I believe that some rights and freedoms need to be limited in the interest of national security, I think that we should end the use of mass surveillance. Moreover, security does not justify brushing aside fundamental right and freedoms.
(predicted <mask> = "While I believe that")
Masked sentence:
I think that we should increase the use of mass surveillance, *<mask>* some rights and freedoms need to be limited in the interest of national security. However, security does not justify brushing aside fundamental rights *and* freedoms.
(gold <mask> = "because")
Output sequence:
I think that we should increase the use of mass surveillance, *and that* some rights and freedoms need to be limited in the interest of national security. However, security does not justify brushing aside fundamental rights *or* freedoms.
(predicted <mask> = "and that")

Table 7

TADM dataset – distribution of samples in the train, dev, and test set according to the discourse-level senses "arg marker" and "disc rel". Columns: "2 ADUs", "3 ADUs", "pred type 1", and "pred type 2" correspond to the selection of samples from the train and dev set by constraining some of the parameters used to generate the TADM dataset (as described in Section 3.4); "all" indicates the total number of samples in TADM dataset, without performing any selection

	all			2 ADUs		3 ADUs		pred type 1		pred type 2	
	test	dev	train	dev	train	dev	train	dev	train	dev	train
arg marker											
backward	120	152	1004	38	251	114	753	0	0	76	502
forward	60	76	502	38	251	38	251	76	502	0	0
rebuttal	180	228	1506	76	502	152	1004	76	502	76	502
thesis	240	304	2008	152	1004	152	1004	152	1004	152	1004
disc rel											
comparison	180	228	1506	76	502	152	1004	76	502	76	502
contingency	120	152	1004	76	502	76	502	76	502	76	502
expansion	60	76	502	0	0	76	502	0	0	0	0
none	240	304	2008	152	1004	152	1004	152	1004	152	1004

Appendix F. TADM dataset – distribution of samples

Table 7 shows the distribution of samples in the train, dev, and test set of TADM dataset according to the discourse-level senses "arg marker" and "disc rel" for the gold DM that should replace the corresponding "<mask>" token. We report the distribution of samples for each of the settings explored in fine-tuning experiments (i.e., "2 ADUs", "3 ADUs", "pred type 1", and "pred type 2", as detailed in Section 3.4). The total number of samples in TADM dataset (without performing any selection of samples) is reported under the "all" column.

Regarding the discourse-level senses "arg marker" and "disc rel", we first observe that the distribution of samples is unbalanced. The selection of gold DMs is based on the DM's role in the templates explored in the TADM dataset. Consequently, these distributions reflect the roles that the DMs play in the TADM dataset. Second, we note the presence of "none" in "disc rel". All these cases correspond to a DM preceding a claim (i.e., "I think that", "in my opinion", or "I believe that", as indicated in Appendix B), which are mapped to the senses "thesis" ("arg marker" sense) and "none" ("disc rel" sense). We recall that for evaluation purposes if the gold DM is mapped to "none", we do not consider the corresponding instance in the evaluation (the DM is out of the scope for the list of DMs available in the senses list, so we cannot make a concrete comparison with the predicted DM).

Appendix G. Fill-in-the-mask discourse marker prediction – error analysis

Table 8 shows some samples and the corresponding gold DMs (accompanied by the discourse-level senses "arg marker" and "disc rel" in parenthesis) from the test set of the TADM dataset. In this table, we highlight some of the challenging samples that can be found in the TADM dataset. More concretely, consecutive pairs of samples in Table 8 contain samples extracted from the TADM dataset that belong to the same instantiation of the core elements but follow different templates, resulting in samples that share some content but also contain targeted edits that require the models to adjust the prediction of DMs accordingly. For instance, samples with id 1 and 2 differ in the parameter "premise role", which changes the premise that is generated and requires the model to predict different DMs

Table 8

Some samples from the test set of the TADM dataset. Columns: "id" corresponds to the sample id, "Masked sentence" to the sample, and "Gold" to the gold DM that should replace the "<mask>" token in the "Masked sentence" (in parenthesis we indicate the discourse-level senses "arg marker" and "disc rel" for the corresponding gold DM)

id	Masked sentence	Gold
1	I think that the use of AI should be abandoned, <mask> these new technologies are not as reliable as conventional ones.	because *(backward)* *(contingency)*
2	I think that the use of AI should be abandoned, <mask> the use of AI is better than the older options.	although *(rebuttal)* *(comparison)*
3	I think that the use of AI should be abandoned, although the use of AI is better than the older options. <mask>, these new technologies are not as reliable as conventional ones.	Moreover *(backward)* *(expansion)*
4	I think that the use of AI should be encouraged, although these new technologies are not as reliable as conventional ones. <mask>, the use of AI is better than the older options.	Moreover *(backward)* *(expansion)*
5	<mask> adolescents are as capable as adults, I think that we should increase youth rights. However, many adolescents can not make responsible decisions.	Because *(forward)* *(contingency)*
6	<mask> adolescents are as capable as adults, I think that we should abolish youth rights. Moreover, many adolescents can not make responsible decisions.	Although *(rebuttal)* *(comparison)*
7	I think that we should increase internet censorship, <mask> enforcement tends to be less effective than persuasion and education. Moreover, a decisive and enforced policy is the best way to deliver a message.	although *(rebuttal)* *(comparison)*
8	I think that we should abandon internet censorship, <mask> enforcement tends to be less effective than persuasion and education. However, a decisive and enforced policy is the best way to deliver a message.	because *(backward)* *(contingency)*

in these samples (despite sharing the same claim); samples 3 and 4 differ in two parameters, "stance role" that dictates the stance of the claim and "supportive premise position" that dictates the position of the supportive premise, requiring the model to predict the same DM in both samples; samples 5 and 6 differ in the parameter "stance role", requiring the model to predict different DMs; and samples 7 and 8 also differ in a single parameter, the "stance role", requiring the model to predict different DMs.

Predictions made by the zero-shot models and fine-tuned `BERT` ("all") model (described in Section 3) for the samples shown in Table 8 are provided in Table 9. As indicated in Section 3.5, we report some confusion between the "arg marker" senses "backward" and "forward" vs. "rebuttal", and the "disc rel" senses "comparison" vs. "expansion", especially on zero-shot models. As previously explained, pairs of samples shown in Table 8 require the models to adjust the prediction of DMs according to the targeted edits performed in different samples. As observed in Table 9, zero-shot `BERT` and `XLM-R` fail to adjust these predictions for all the pairs of samples presented in Table 8, keeping the same prediction (or at least the same sense) for different samples that belong to the same instantiation of the core elements, evidence that these models are not robust to the targeted edits. These erroneous predictions result

Table 9

Zero-shot and fine-tuned models (described in Section 3) predictions for some samples extracted from the TADM dataset. "id": sample id (corresponds to the same sample id from Table 8). For each prediction, we provide the predicted DM and the corresponding discourse-level senses ("arg marker" and "disc rel") in parenthesis

id	Zero-shot				Fine-tuned
	BERT	XLM-R	BART V1	BART V2	BERT (all)
1	as	because	because	because	because
	(none)	*(backward)*	*(backward)*	*(backward)*	*(backward)*
	(temporal)	*(contingency)*	*(contingency)*	*(contingency)*	*(contingency)*
2	because	because	but	but I think that it is possible that	although
	(backward)	*(backward)*	*(rebuttal)*	*(none)*	*(rebuttal)*
	(contingency)	*(contingency)*	*(comparison)*	*(none)*	*(comparison)*
3	Unfortunately	However	However	However	Moreover
	(none)	*(rebuttal)*	*(rebuttal)*	*(rebuttal)*	*(backward)*
	(none)	*(comparison)*	*(comparison)*	*(comparison)*	*(expansion)*
4	However	However	However	However, I think that in the long run, in the short term	Moreover
	(rebuttal)	*(rebuttal)*	*(rebuttal)*	*(none)*	*(backward)*
	(comparison)	*(comparison)*	*(comparison)*	*(none)*	*(expansion)*
5	If	Because	Since	Since	Because
	(none)	*(forward)*	*(backward)*	*(backward)*	*(forward)*
	(contingency)	*(contingency)*	*(contingency)*	*(contingency)*	*(contingency)*
6	If	Because	Since	Since	Although
	(none)	*(forward)*	*(backward)*	*(backward)*	*(rebuttal)*
	(contingency)	*(contingency)*	*(contingency)*	*(contingency)*	*(comparison)*
7	since	because	but	but I think that	although
	(backward)	*(backward)*	*(rebuttal)*	*(none)*	*(rebuttal)*
	(contingency)	*(contingency)*	*(comparison)*	*(none)*	*(comparison)*
8	because	because	as	as	because
	(backward)	*(backward)*	*(none)*	*(none)*	*(backward)*
	(contingency)	*(contingency)*	*(temporal)*	*(temporal)*	*(contingency)*

in some confusion between discourse-level senses, namely "backward" vs. "rebuttal" (samples 2, 3, 4, and 7), "forward" vs. "rebuttal" (sample 6), "comparison" vs. "contingency" (samples 2, 6, and 7), and "comparison" vs "expansion" (samples 3 and 4). As reported in Section 3.5, we observed less confusion between these labels with zero-shot BART-based models. The sample pair 1 and 2 in Table 9 illustrates this point: BART-based models adjust the predictions of DMs accordingly. However, some of the confusions previously reported for zero-shot `BERT` and `XLM-R` are also observed for zero-shot BART-based models, as exemplified with samples 3, 4, and 6 (Table 9). Finally, as indicated in Section 3.5, fine-tuned `BERT` ("all") performs better than zero-shot models. These improvements can be observed in Table 9, where fine-tuned `BERT` ("all") adjusts the predictions of DMs in the correct direction for all the samples (i.e., predicted DM discourse-level senses match the gold data). These results are evidence that after some fine-tuning, LMs can improve their robustness to the challenging samples contained in the TADM dataset.

Table 10

Some samples used in human evaluation experiments. Columns: "id" corresponds to the sample id, "Masked sentence" to the sample extracted from the TADM dataset, "Prediction" to the DM prediction made by one of the models analyzed in a zero-shot setting (as described in Section 3.3), "Grammaticality" ("Gram.") and "Coherence" ("Coh.") to the human assessment criteria (Section 3.6)

id	Masked sentence	Prediction	Gram.	Coh.
1	I think that we should abolish environmental laws, <mask> environmentalism stands in the way of technological progress and economic growth.	because	(+1)	(+1)
2	I think that we should abolish environmental laws, <mask> people must protect nature and respect its biological communities.	but	(+1)	(+1)
3	I think that we should abolish environmental laws, <mask> people must protect nature and respect its biological communities.	because	(+1)	(−1)
4	I think that we should introduce goal line technology, <mask> the current system is working, and making such a change could have negative consequences. Moreover, it is time to change the old ways and try something new.	because	(+1)	(−1)
5	Although it is time to change the old ways and try something new, I think that we should oppose goal line technology. <mask>, the current system is working, and making such a change could have negative consequences.	However	(+1)	(−1)
6	Although it is time to change the old ways and try something new, I think that we should oppose goal line technology. <mask>, the current system is working, and making such a change could have negative consequences.	In	(−1)	(0)
7	<mask> animals should not be treated as property, I think that bullfighting should be legalized.	While	(+1)	(+1)
8	<mask> animals should not be treated as property, I think that bullfighting should be legalized.	I think that	(0)	(−1)
9	<mask> animals should not be treated as property, I think that bullfighting should be legalized.	Because	(+1)	(−1)
10	<mask> animals should not be treated as property, I think that bullfighting should be legalized.	If	(+1)	(−1)
11	I think that big governments should be abandoned, <mask> public utility is best served by actions coordinated by central government.	because	(+1)	(−1)
12	I think that big governments should be abandoned, <mask> public utility is best served by actions coordinated by central government.	and	(+1)	(−1)
13	I think that big governments should be abandoned, <mask> public utility is best served by actions coordinated by central government.	and that	(+1)	(−1)
14	<mask> people who come in search of a safer and better life should not be turned away, I think that we should protect the right of asylum. However, mass immigration threatens social cohesion.	I	(−1)	(0)
15	<mask> we should abolish anti-social behavior orders, because strict punishment is not effective in preventing criminal behavior. However, when people will have to pay for their actions there will be less crime.	So	(+1)	(+1)

Appendix H. Human evaluation – additional details

Table 10 shows some of the samples extracted from the TADM dataset that were analyzed in human evaluation experiments. For each sample, we provided to the annotators the "Masked sentence" received as input by the zero-shot models explored in the fill-in-the-mask discourse marker prediction task (Section 3) and the corresponding "Prediction" made by the models. As detailed in Section 3.6, each annotator rated the samples based on the criteria "Grammaticality" and "Coherence". The ratings presented in Table 10 correspond to the majority voting obtained from the ratings provided by the three annotators.

Table 11

Human evaluation – confusion matrices between each pair of annotators ("A1", "A2", and "A3") for the ratings (i.e., (−1), (0), or (+1)) provided in the "Grammaticality" criterion (as described in Section 3.6). Numbers correspond to raw counts

A1/A2	(−1)	(0)	(+1)
(−1)	12	0	0
(0)	1	0	2
(+1)	0	2	50

A1/A3	(−1)	(0)	(+1)
(−1)	12	0	0
(0)	0	0	3
(+1)	1	5	46

A2/A3	(−1)	(0)	(+1)
(−1)	12	0	1
(0)	0	2	0
(+1)	1	3	48

Table 12

Human evaluation – confusion matrices between each pair of annotators ("A1", "A2", and "A3") for the ratings (i.e., (−1), (0), or (+1)) provided in the "Coherence" criterion (as described in Section 3.6). Numbers correspond to raw counts

A1/A2	(−1)	(0)	(+1)
(−1)	17	3	1
(0)	0	7	4
(+1)	2	5	28

A1/A3	(−1)	(0)	(+1)
(−1)	12	6	3
(0)	0	8	3
(+1)	0	10	25

A2/A3	(−1)	(0)	(+1)
(−1)	11	5	3
(0)	1	13	1
(+1)	0	6	27

Tables 11 and 12 show the confusion matrices between each pair of annotators ("A1", "A2", and "A3") for the ratings provided in the "Grammaticality" and "Coherence' criteria, respectively.

Regarding the "Grammaticality" criterion, most samples (approximately 76%) are rated as (+1). Annotators rated (−1) and (0) on 19% and 5% of the samples, respectively. These percentages correspond to the average percentage over all pairs of annotators (based on Table 11). Indeed, most of the predictions made by the models correspond to conventional and sound DMs in terms of "Grammaticality" (as exemplified in the set of samples shown in Table 10). However, in some cases, the annotators consider the predictions ungrammatical. These predictions rated as ungrammatical occur when the "<mask>" token is placed at the beginning of a sentence, with the models predicting prepositions or personal pronouns in those cases (e.g., "In" in sample 6 and "I" in sample 14 from Table 10). We attribute such erroneous predictions to samples that the models were not able to capture the meaning of the sentence, resorting to their language modeling capabilities by predicting the most likely sequence to replace the "<mask>" token (prepositions and personal pronouns are commonly used at the beginning of sentences). Analyzing the disagreements between annotators (Table 11), we observe that most of them occur when one of the annotators provides the rating (0) while the other annotator considers either (+1) or (−1), which corresponds to 84% of the disagreements. We highlight two recurring cases: (a) when the model predicts a DM already in the sentence (as illustrated in sample 8 from Table 10): in these cases, some annotators considered that repetitions are not appropriate in this axis, rating these cases as (0), while other annotators consider appropriate (+1) if aligned with the argument positioning; and (b) some annotators made a distinction between the predictions "and" vs. "and that" (e.g., rating sample 12 from Table 10) with (0) and sample 13 with (+1)), while other annotators rated both predictions as appropriate (+1).

Regarding "Coherence", 49% of the samples are rated as (+1), while the percentage of samples rated with (−1) and (0) is similar (26% and 25%, respectively). The lower percentage of (+1) in "Coherence" compared to "Grammaticality" indicates that the models struggle in this axis (i.e., to predict a DM that conveys a correct connotation given the context, especially for the challenging "edge cases" in the TADM dataset, as detailed in Section 3.5). Indeed, we observe that most predictions rated as (−1) correspond to DM predictions not aligned with the connotation of the ADU they precede. Analyzing the confusing matrix in Table 12, we observe that most of the disagreements occur when one of the annotators provides the rating (0) while the other annotator considers either (+1) or (−1), which corresponds to 83% of the

Table 13

End-to-end DM augmentation results – small-scale experiment. Columns: "seq2seq model" indicates the pre-trained Seq2Seq LM, "fine-tune data" indicates the data used to fine-tune the Seq2Seq LM (details in Section 5.2), the metrics under "explicit DMs accuracy analysis" correspond to the automatic evaluation metrics (Section 3.2), the scores in the "coverage analysis" column correspond to the percentage of DMs that are predicted (as detailed in Section 5.1)

seq2seq model	fine-tune data	explicit DMs accuracy analysis					coverage analysis
		word embs	retrofit embs	sbert embs	arg marker	disc rel	pct. DMs predicted
ChatGPT	none	0.3630	0.1972	0.2056	0.2222	**0.3600**	0.5476
BART-base	none	0	0	0	0	0	0.0163
	Discovery	0.5269	0.1965	0.2503	0.3819	0.1800	**0.9443**
	TADM	**0.5807**	**0.2635**	0.2697	0.2639	0.2800	0.8626
	PDTB	0.4009	0.2021	0.1953	0.1875	0.3200	0.6582
BART-base	Discovery +	0.5400	0.2326	0.2937	0.4097	0.2400	0.9274
T5-base	TADM + PDTB	0.5367	0.2420	0.2973	0.4861	0.3000	0.9200
T5-large		0.5282	0.2548	**0.2974**	**0.5139**	0.3400	0.8978

disagreements. Regarding these disagreements, we make the following observations: (a) some samples seem to require domain knowledge about the topic that is relevant to understanding the positioning of ADUs, which was not clear for all annotators. For instance, consider sample 11 from Table 10. One annotator rated "Coherence" with (+1), while the others rated it with (−1). In this context, "central government" should be considered as a "big government", hence (−1) is the most appropriate rating; (b) when the mask is positioned at the beginning of the sentence and the model predicts DMs such as "Therefore" or "So" (e.g., sample 15 from Table 10), some annotators considered that it demands more context, while other annotators were agnostic to it, leading to some disagreements; (c) some predictions can be considered as "neutral" in terms of positioning, leading to different interpretations of ADU roles in the argument (e.g., prediction of "If" in sample 10, one annotator rated "Coherence" with (0) and the others with (−1)).

Appendix I. End-to-end DM augmentation results – comparison with ChatGPT

Table 13 shows the results obtained for the small-scale end-to-end DM augmentation experiment with ChatGPT, including both the explicit DMs accuracy and coverage analysis.

Appendix J. Annotation projection

To map the label sequence from the original sequence to the modified sequence, we implement the Needleman-Wunsch algorithm [42], a well-known sequence alignment algorithm. As input, it receives the original and modified token sequences. The output is an alignment of the token sequences, token by token, where the goal is to optimize a global score. This algorithm might include a special token (the "gap" token) in the output sequences. Gap tokens are inserted to optimize the alignment of identical tokens in successive sequences. The global score attributed to a given alignment is based on a scoring system. We use default values: match score (tokens are identical) = 1, mismatch score (tokens are different but aligned to optimize alignment sequence) = −1, gap penalty (gap token was introduced in one of the sequences) = −1. To determine whether two tokens are identical, we use strict lexical match (case insensitive). Using the aligned sequences, we map the labels from the original to the modified token sequence.

Table 14

Annotation projection examples for each ArgMining corpora. "Original text": original text and gold annotations for ADU identification (represented using square brackets) and classification (acronyms in subscript). "Text augmented with DMs + Annotation Projection": text obtained after performing DM augmentation (using T5-base fine-tuned on "Discovery + TADM + PDTB" when we provide "Input data: removed DMs") and the labels obtained after annotation projection. The underlined text highlights differences between original and DM augmented text

PEC
Original text:
(. . .) Considering this fact, [advertisements have undeniable affects on the society about the product being advertised]$_P$. [They make the product preferable]$_C$.
Text augmented with DMs + Annotation Projection:
(. . .) *Secondly,* [advertisements have undeniable affects on the society about the product being advertised]$_P$. *Hence,* [they make the product preferable]$_C$.
MTX
Original text:
[Alternative treatments should be subsidized in the same way as conventional treatments,]$_C$ [since both methods can lead to the prevention, mitigation or cure of an illness.]$_P$ (. . .)
Text augmented with DMs + Annotation Projection:
[*I think that* alternative treatments should be subsidized in the same way as conventional treatments,]$_C$ [*because* both methods can lead to the prevention, mitigation or cure of an illness.]$_P$ (. . .)
Hotel
Original text:
(. . .) [The shower was a bit short for my tall husband.]$_C$ [The beds were fabulous!]$_C$ [I loved the body pillow too.]$_C$ [We would definitely stay here again.]$_R$
Text augmented with DMs + Annotation Projection:
(. . .) [*Although,* the shower was a bit short for my tall husband.]$_C$ [*Otherwise,* the beds were fabulous!]$_C$ [*And,* I loved the body pillow too.]$_C$ [*Overall,* we would definitely stay here again.]$_R$

Table 14 illustrates some examples of the output obtained when employing the annotation projection procedure described in this section to the three ArgMining corpora explored in this work. "Original text" corresponds to the original text from which gold annotations for ADU identification and classification are provided in the ArgMining corpora. "Text augmented with DMs + Annotation Projection" corresponds to the text obtained after performing DM augmentation (using the pre-trained Seq2Seq model `T5-base` fine-tuned on the combination of the corpora "Discovery + TADM + PDTB", as described in Section 6.2) when we provide as input the text deprived of DMs (i.e., "Input data: removed DMs") and the corresponding ADU identification and classification labels obtained after performing annotation projection. The underlined text highlights differences between the original and DM augmented text. These differences require a projection of the original label sequence to the label sequence for the corresponding DM augmented text, which is performed using the proposed annotation projection procedure.

Appendix K. Downstream task evaluation – error analysis

We show some examples of the gold data and predictions made by the downstream task models for the ArgMining corpora explored in this work.

For each example, we provide: (a) "Gold": the gold data, including the ADU boundaries (square brackets) and the ADU labels (acronyms in subscript); and (b) "Input data: X (Y)": where "X" indicates

Table 15

Downstream task – PEC example containing C and MC in the gold annotation. "Gold": original text and gold annotations for ADU identification (represented using square brackets) and classification (acronyms in subscript). "Input data: X (Y)": predictions made by a BERT model on four different experimental setups, where "X" indicates the version of the input data provided to the DM augmentation model, and "Y" indicates whether DM augmentation is performed. The underlined text highlights the presence of DMs

Gold
In a word, [our parents and other family members can help us a lot in our life]$_C$. *I believe that* [with the help of our family, we can achieve success quite easily]$_{MC}$.
Input data: removed DMs (none)
[Our parents and other family members can help us a lot in our life]$_C$. [With the help of our family, we can achieve success quite easily]$_C$.
Input data: removed DMs (T5-base)
However, [our parents and other family members can help us a lot in our life]$_{MC}$. *Hence,* [with the help of our family, we can achieve success quite easily]$_C$.
Input data: original (none)
In a word, [our parents and other family members can help us a lot in our life]$_C$. *I believe that* [with the help of our family, we can achieve success quite easily]$_{MC}$.
Input data: original (T5-base)
So in a word, [our parents and other family members can help us a lot in our life]$_{MC}$. *Personally, I believe that* [with the help of our family, we can achieve success quite easily]$_{MC}$.

the version of the input data provided to the DM augmentation model, and "Y" indicates whether we perform DM augmentation or not ("none" indicates that we do not perform DM augmentation and `T5-base` indicates that we perform DM augmentation using the pre-trained Seq2Seq model `T5-base` fine-tuned on the combination of the corpora "Discovery + TADM + PDTB").

Tables 15 and 16 show two examples from PEC. Table 15 shows a paragraph containing a C and MC in the "Gold" data annotations. We observe that in the "Input data: original (none)" setup, the model predicts MC frequently in the presence of a DM that can be mapped to the "arg marker" sense "thesis" (e.g., "in conclusion", "in my opinion", "as far as I am concerned", "I believe that", etc.). Similar patterns can be observed in the "Gold" data annotations. We were not able to find similar associations in the "Input data: removed DMs (`T5-base`)" setup, for instance. As illustrated in "Input data: removed DMs (none)", the distinction between C and MC is very challenging in the absence of such explicit signals. The distinction between C and P can also be challenging, as exemplified in Table 16. We observe that some DMs might be associated to ADU labels more strongly than others (e.g., in Table 16, "therefore" is associated to C predictions, while "firstly" cannot be associated to a particular label). Surprisingly, we observed that some DMs that are commonly associated as indicators of either C or P ADUs (e.g., "because" and "moreover" typically associated to P) are not consistently used by the downstream model to predict the corresponding ADU label accordingly.

Table 17 shows an example from MTX. Regarding the setups containing the original data (i.e., "Gold" annotations and the predictions made for "Input data: original (none)"), besides a single occurrence of "therefore" and "nevertheless", all the remaining C do not contain a DM preceding them (this analysis is constrained to the test set). Some of the ADUs labeled as P are preceded with DMs (most common DMs are: "and" (6), "but" (10), "yet" (4), and "besides" (3)), even though most of them (44) are not preceded by a DM (numbers in parentheses correspond to the number of occurrences in the test set for the "Gold" annotations, similar numbers are obtained for "Input data: original (none)"). DM augmentation approaches performed well in terms of coverage, with most of the ADUs being preceded by DMs. We can observe in Table 17 that some ADU labels become more evident after the DM augmentation

Table 16

Downstream task – PEC example containing P and C in the gold annotation. "Gold": original text and gold annotations for ADU identification (represented using square brackets) and classification (acronyms in subscript). "Input data: X (Y)": predictions made by a BERT model on four different experimental setups, where "X" indicates the version of the input data provided to the DM augmentation model, and "Y" indicates whether DM augmentation is performed. The underlined text highlights the presence of DMs

Gold
[We are mainly introduced to products through advertisements]$_P$. *Therefore,* [advertisers push the limits of creativity to dispose the consumers to purchase the product]$_C$. [When the consumers are impressed by the way a product is advertised, they can be convinced to consider that the product is a need in some cases]$_P$. [Recently, there is a very creative advertisements of a soft drink product on TV]$_P$. [The story delivers a desired call to drink that soft drink that people tend to drink when the weather is too hot]$_P$. *As a result of this,* [the number of that product being sold will increases]$_P$.
Input data: removed DMs (none)
[We are mainly introduced to products through advertisements]$_C$. [Advertisers push the limits of creativity to dispose the consumers to purchase the product]$_P$. [When the consumers are impressed by the way a product is advertised, they can be convinced to consider that the product is a need in some cases]$_P$. [Recently, there is a very creative advertisements of a soft drink product on TV]$_P$. [The story delivers a desired call to drink that soft drink that people tend to drink when the weather is too hot]$_P$. [The number of that product being sold will increases]$_C$.
Input data: removed DMs (T5-base)
Firstly, [we are mainly introduced to products through advertisements]$_C$. *Secondly,* [advertisers push the limits of creativity to dispose the consumers to purchase the product]$_P$. *Secondly,* [when the consumers are impressed by the way a product is advertised, they can be convinced to consider that the product is a need in some cases]$_P$. *For example,* [recently, there is a very creative advertisements of a soft drink product on TV]$_P$. [The story delivers a desired call to drink that soft drink that people tend to drink when the weather is too hot]$_P$. *Thus,* [the number of that product being sold will increases]$_P$.
Input data: original (none)
[We are mainly introduced to products through advertisements]$_P$. *Therefore,* [advertisers push the limits of creativity to dispose the consumers to purchase the product]$_C$. [When the consumers are impressed by the way a product is advertised, they can be convinced to consider that the product is a need in some cases]$_P$. [Recently, there is a very creative advertisements of a soft drink product on TV]$_P$. [The story delivers a desired call to drink that soft drink that people tend to drink when the weather is too hot]$_P$. *As a result of this,* [the number of that product being sold will increases]$_P$.
Input data: original (T5-base)
Firstly, [we are mainly introduced to products through advertisements]$_P$. *Therefore,* [advertisers push the limits of creativity to dispose the consumers to purchase the product]$_C$. *Secondly,* [when the consumers are impressed by the way a product is advertised, they can be convinced to consider that the product is a need in some cases]$_P$. *For example,* [recently, there is a very creative advertisements of a soft drink product on TV]$_P$. [The story delivers a desired call to drink that soft drink that people tend to drink when the weather is too hot]$_P$. *As a result of this,* [the number of that product being sold will increases]$_P$.

performed by the models proposed in this work ("Input data: removed DMs (`T5-base`)" and "Input data: original (`T5-base`)"), such as the presence of the DM "clearly" indicating C and the presence of "besides", "because" or "but" indicating P.

Finally, Table 18 shows an example from Hotel. Similar to the observations made for MTX, in the setups containing the original data (i.e., "Gold" annotations and the predictions made for "Input data: original (none)"), most ADUs are not preceded by DMs. The only exception is the DM "and" that occurs with some frequency preceding C (10 out of 199 ADUs labeled as C) and P (4 out of 41). For instance, in Table 18, 9 ADUs were annotated and none of them is preceded by a DM; making the annotation of ADUs (arguably) very challenging. Despite the lack of explicit clues, downstream models perform relatively well in this example, only missing the two gold IPs (not identified as an ADU in one of the cases and predicted as C in the other case) and erroneously labeling as C the only sentence in the gold data that is not annotated as an ADU. Also similar to MTX, DM augmentation approaches performed well in terms of coverage, with most ADUs being preceded by DMs. However, as observed in Table 18,

Table 17

Downstream task – MTX example. "Gold": original text and gold annotations for ADU identification (represented using square brackets) and classification (acronyms in subscript). "Input data: X (Y)": predictions made by a BERT model on four different experimental setups, where "X" indicates the version of the input data provided to the DM augmentation model, and "Y" indicates whether DM augmentation is performed. The underlined text highlights the presence of DMs

Gold
[One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt.]$_P$ [*And* when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles.]$_P$ [Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners.]$_C$ [*Of course,* first they'd actually need to be caught in the act by public order officers,]$_P$ [*but* once they have to dig into their pockets, their laziness will sure vanish!]$_P$
Input data: removed DMs (none)
[One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt.]$_P$ [When bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles.]$_P$ [Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners.]$_C$ [First they'd actually need to be caught in the act by public order officers,]$_P$ [once they have to dig into their pockets, their laziness will sure vanish!]$_P$
Input data: removed DMs (T5-base)
[*But* one can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt.]$_P$ [*Besides,* when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles.]$_P$ [*Clearly,* higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners.]$_C$ [*Because,* first they'd actually need to be caught in the act by public order officers,]$_P$ [*but* once they have to dig into their pockets, their laziness will sure vanish!]$_P$
Input data: original (none)
[One can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt.]$_P$ [*And* when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles.]$_P$ [Higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners.]$_C$ [*Of course,* first they'd actually need to be caught in the act by public order officers,]$_P$ [*but* once they have to dig into their pockets, their laziness will sure vanish!]$_P$
Input data: original (T5-base)
[*But* one can hardly move in Friedrichshain or Neukölln these days without permanently scanning the ground for dog dirt.]$_P$ [*Besides, and* when bad luck does strike and you step into one of the many 'land mines' you have to painstakingly scrape the remains off your soles.]$_P$ [*Clearly,* higher fines are therefore the right measure against negligent, lazy or simply thoughtless dog owners.]$_C$ [*Of course,* first they'd actually need to be caught in the act by public order officers,]$_P$ [*but* once they have to dig into their pockets, their laziness will sure vanish!]$_P$

the impact on the downstream model predictions is small (the predictions for all the setups are similar, the only exception is the extra split on "so was the bathroom" performed in "Input data: removed DMs (`T5-base`)", even though this span of text is similar in all setups). We point out that, particularly in this text genre, adding DMs to signal the presence of ADUs might contribute to improving the readability of arguments exposed in the text, as exemplified by the DM augmentation performed by the models proposed in this work ("Input data: removed DMs (`T5-base`)" and "Input data: original (`T5-base`)" in Table 18).

Table 18

Downstream task – Hotel example. "Gold": original text and gold annotations for ADU identification (represented using square brackets) and classification (acronyms in subscript). "Input data: X (Y)": predictions made by a BERT model on four different experimental setups, where "X" indicates the version of the input data provided to the DM augmentation model, and "Y" indicates whether DM augmentation is performed. The underlined text highlights the presence of DMs

Gold

[Great Hotel for Vegas First Timers]$_{MC}$ [We stayed at the MGM from Jan. 31 to Feb. 3 for UFC 94.]$_{B}$ [We arrived at 11 am but were able to check in early.]$_{IP}$ [We had a view of mountains and palm trees but no strip.]$_{IP}$ We didn't spend much time in our room anyway. [The rooms were large, so was the bathroom.]$_{C}$ [The shower was a bit short for my tall husband.]$_{C}$ [The beds were fabulous!]$_{C}$ [I loved the body pillow too.]$_{C}$ [We would definitely stay here again.]$_{R}$

Input data: removed DMs (none)

[Great Hotel for Vegas First Timers]$_{MC}$ [We stayed at the MGM from Jan. 31 to Feb. 3 for UFC 94.]$_{B}$ We arrived at 11 am but were able to check in early. [We had a view of mountains and palm trees but no strip.]$_{C}$ [We didn't spend much time in our room anyway.]$_{C}$ [The rooms were large, so was the bathroom.]$_{C}$ [The shower was a bit short for my tall husband.]$_{C}$ [The beds were fabulous!]$_{C}$ [I loved the body pillow too.]$_{C}$ [We would definitely stay here again.]$_{R}$

Input data: removed DMs (T5-base)

[Great Hotel for Vegas First Timers]$_{MC}$ [We stayed at the MGM from Jan. 31 to Feb. 3 for UFC 94.]$_{B}$ *Specifically,* we arrived at 11 am but were able to check in early. [*Besides,* we had a view of mountains and palm trees but no strip.]$_{C}$ [*So,* we didn't spend much time in our room anyway.]$_{C}$ [*Although,* the rooms were large,]$_{C}$ [so was the bathroom.]$_{C}$ [*Although,* the shower was a bit short for my tall husband.]$_{C}$ [*Otherwise,* the beds were fabulous!]$_{C}$ [*And,* I loved the body pillow too.]$_{C}$ [*Overall,* we would definitely stay here again.]$_{R}$

Input data: original (none)

[Great Hotel for Vegas First Timers]$_{MC}$ [We stayed at the MGM from Jan. 31 to Feb. 3 for UFC 94.]$_{B}$ We arrived at 11 am but were able to check in early. [We had a view of mountains and palm trees but no strip.]$_{C}$ [We didn't spend much time in our room anyway.]$_{C}$ [The rooms were large, so was the bathroom.]$_{C}$ [The shower was a bit short for my tall husband.]$_{C}$ [The beds were fabulous!]$_{C}$ [I loved the body pillow too.]$_{C}$ [We would definitely stay here again.]$_{R}$

Input data: original (T5-base)

[Great Hotel for Vegas First Timers]$_{MC}$ [We stayed at the MGM from Jan. 31 to Feb. 3 for UFC 94.]$_{B}$ *Specifically,* we arrived at 11 am but were able to check in early. [*Besides,* we had a view of mountains and palm trees but no strip.]$_{C}$ [*So,* we didn't spend much time in our room anyway.]$_{C}$ [*Although,* the rooms were large, so was the bathroom.]$_{C}$ [*Although,* the shower was a bit short for my tall husband.]$_{C}$ [*Otherwise,* the beds were fabulous!]$_{C}$ [*And,* I loved the body pillow too.]$_{C}$ [*Overall,* we would definitely stay here again.]$_{R}$

References

[1] P. Accuosto and H. Saggion, Transferring knowledge from discourse to arguments: A case study with scientific abstracts, in: *Proceedings of the 6th Workshop on Argument Mining*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 41–51, https://aclanthology.org/W19-4505. doi:10.18653/v1/W19-4505.

[2] J. Bao, C. Fan, J. Wu, Y. Dang, J. Du and R. Xu, A neural transition-based model for argumentation mining, in: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Association for Computational Linguistics, 2021, pp. 6354–6364, Online, https://aclanthology.org/2021.acl-long.497. doi:10.18653/v1/2021.acl-long.497.

[3] I. Beltagy, M.E. Peters and A. Cohan, *Longformer: The Long-Document Transformer, CoRR* (2020), https://arxiv.org/abs/2004.05150 arXiv:2004.05150.

[4] Y. Bilu, A. Gera, D. Hershcovich, B. Sznajder, D. Lahav, G. Moshkowich, A. Malet, A. Gavron and N. Slonim, Argument invention from first principles, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1013–1026, https://aclanthology.org/P19-1097. doi:10.18653/v1/P19-1097.

[5] O. Biran and O. Rambow, Identifying justifications in written dialogs by classifying text as argumentative, *Int. J. Semantic Comput.* **5**(4) (2011), 363–381. doi:10.1142/S1793351X11001328.

[6] C. Braud and P. Denis, Combining natural and artificial examples to improve implicit discourse relation identification, in: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014, pp. 1694–1705, https://aclanthology.org/C14-1160.

[7] C. Braud and P. Denis, Learning connective-based word representations for implicit discourse relation identification, in: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Austin, Texas, 2016, pp. 203–213, https://aclanthology.org/D16-1020. doi:10.18653/v1/D16-1020.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever and D. Amodei, Language models are few-shot learners, in: *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin, eds, Vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901, https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[9] A. Chernodub, O. Oliynyk, P. Heidenreich, A. Bondarenko, M. Hagen, C. Biemann and A. Panchenko, TARGER: Neural argument mining at your fingertips, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 195–200, https://aclanthology.org/P19-3031. doi:10.18653/v1/P19-3031.

[10] O. Cocarascu, E. Cabrio, S. Villata and F. Toni, Dataset independent baselines for relation prediction in argument mining, in: *Computational Models of Argument – Proceedings of COMMA 2020*, Perugia, Italy, September 4–11, 2020, H. Prakken, S. Bistarelli, F. Santini and C. Taticchi, eds, Frontiers in Artificial Intelligence and Applications, Vol. 326, IOS Press, 2020, pp. 45–52. doi:10.3233/FAIA200490.

[11] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer and V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 8440–8451, Online, https://aclanthology.org/2020.acl-main.747. doi:10.18653/v1/2020.acl-main.747.

[12] D. Das, T. Scheffler, P. Bourgonje and M. Stede, Constructing a lexicon of English discourse connectives, in: *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 360–365, https://aclanthology.org/W18-5042. doi:10.18653/v1/W18-5042.

[13] D. Das and M. Taboada, RST signalling corpus: A corpus of signals of coherence relations, *Lang. Resour. Eval.* **52**(1) (2018), 149–184. doi:10.1007/s10579-017-9383-x.

[14] M. Davies and J.L. Fleiss, Measuring agreement for multinomial data, *Biometrics* (1982), 1047–1051. doi:10.2307/2529886.

[15] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] S. Dutta, J. Juneja, D. Das and T. Chakraborty, Can unsupervised knowledge transfer from social discussions help argument mining? in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 7774–7786, https://aclanthology.org/2022.acl-long.536. doi:10.18653/v1/2022.acl-long.536.

[17] S. Eger, J. Daxenberger and I. Gurevych, Neural end-to-end learning for computational argumentation mining, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 11–22, https://aclanthology.org/P17-1002. doi:10.18653/v1/P17-1002.

[18] S. Eger, J. Daxenberger, C. Stab and I. Gurevych, Cross-lingual argumentation mining: Machine translation (and a bit of projection) is all you need! in: *Proceedings of the 27th International Conference on Computational Linguistics*, Association for Computational Linguistics, Santa Fe, New Mexico, USA, 2018, pp. 831–844, https://aclanthology.org/C18-1071.

[19] S. Eger, A. Rücklé and I. Gurevych, PD3: Better low-resource cross-lingual transfer by combining direct transfer and annotation projection, in: *Proceedings of the 5th Workshop on Argument Mining*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 131–143, https://aclanthology.org/W18-5216. doi:10.18653/v1/W18-5216.

[20] B. Fraser, What are discourse markers?, *Journal of Pragmatics* **31**(7) (1999), 931–952. doi:10.1016/S0378-2166(98)00101-5.

[21] B. Fraser, An account of discourse markers, *International Review of Pragmatics* **1**(2) (2009), 293–320. doi:10.1163/187730909X12538045489818.

[22] A. Galassi, M. Lippi and P. Torroni, Multi-Task Attentive Residual Networks for Argument Mining, 2021, CoRR, https://arxiv.org/abs/2102.12227 arXiv:2102.12227.

[23] Y. Gao, H. Wang, C. Zhang and W. Wang, Reinforcement Learning Based Argument Component Detection, 2017, CoRR, http://arxiv.org/abs/1702.06239 arXiv:1702.06239.

[24] I. Habernal, D. Faber, N. Recchia, S. Bretthauer, I. Gurevych, I. Spiecker genannt Döhmann and C. Burchard, Mining Legal Arguments in Court Decisions, 2022, arXiv preprint. doi:10.48550/arXiv.2208.06178.

[25] F. Hewett, R. Prakash Rane, N. Harlacher and M. Stede, The utility of discourse parsing features for predicting argumentation structure, in: *Proceedings of the 6th Workshop on Argument Mining*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 98–103, https://aclanthology.org/W19-4512. doi:10.18653/v1/W19-4512.

[26] H.-P. Huang and J.J. Li, Unsupervised adversarial domain adaptation for implicit discourse relation classification, in: *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 686–695, https://aclanthology.org/K19-1064. doi:10.18653/v1/K19-1064.

[27] L. Huber, Y. Toussaint, C. Roze, M. Dargnat and C. Braud, Aligning discourse and argumentation structures using subtrees and redescription mining, in: *Proceedings of the 6th Workshop on Argument Mining*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 35–40, https://aclanthology.org/W19-4504. doi:10.18653/v1/W19-4504.

[28] C. Jiang, T. Qian, Z. Chen, K. Tang, S. Zhan and T. Zhan, Generating pseudo connectives with MLMs for implicit discourse relation recognition, in: *PRICAI 2021: Trends in Artificial Intelligence*, D.N. Pham, T. Theeramunkong, G. Governatori and F. Liu, eds, Springer, Cham, 2021, pp. 113–126. ISBN 978-3-030-89363-7.

[29] M. Kurfalı and R. Östling, Let's be explicit about that: Distant supervision for implicit discourse relation classification via connective prediction, in: *Proceedings of the 1st Workshop on Understanding Implicit and Underspecified Language*, Association for Computational Linguistics, 2021, pp. 1–10, Online, https://aclanthology.org/2021.unimplicit-1.1. doi:10.18653/v1/2021.unimplicit-1.1.

[30] T. Kuribayashi, H. Ouchi, N. Inoue, P. Reisert, T. Miyoshi, J. Suzuki and K. Inui, An empirical study of span representations in argumentation structure parsing, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4691–4698, https://aclanthology.org/P19-1464. doi:10.18653/v1/P19-1464.

[31] M. Lan, J. Wang, Y. Wu, Z.-Y. Niu and H. Wang, Multi-task attention-based neural networks for implicit discourse relationship representation and identification, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1299–1308, https://aclanthology.org/D17-1134. doi:10.18653/v1/D17-1134.

[32] J.R. Landis and G.G. Koch, The measurement of observer agreement for categorical data, *Biometrics* **33**(1) (1977), 159–174, http://www.jstor.org/stable/2529310. doi:10.2307/2529310.

[33] J. Lawrence and C. Reed, Argument mining: A survey, *Computational Linguistics* **45**(4) (2019), 765–818, https://aclanthology.org/J19-4006. doi:10.1162/coli_a_00364.

[34] C. Leiter, R. Zhang, Y. Chen, J. Belouadi, D. Larionov, V. Fresen and S. Eger, ChatGPT: A Meta-Analysis after 2.5 Months, (2023), *ArXiv* arXiv:2302.13795.

[35] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov and L. Zettlemoyer, BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, 2020, pp. 7871–7880, Online, https://aclanthology.org/2020.acl-main.703. doi:10.18653/v1/2020.acl-main.703.

[36] M. Lippi and P. Torroni, Argumentation Mining: State of the Art and Emerging Trends, *ACM Trans. Internet Technol.* **16**(2) (2016). doi:10.1145/2850417.

[37] H. Liu, Y. Gao, P. Lv, M. Li, S. Geng, M. Li and H. Wang, Using argument-based features to predict and analyse review helpfulness, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Copenhagen, Denmark, 2017, pp. 1358–1363, https://aclanthology.org/D17-1142. doi:10.18653/v1/D17-1142.

[38] W.C. Mann and S.A. Thompson, Rhetorical structure theory: Description and construction of text structures, in: *Natural Language Generation: New Results in Artificial Intelligence, Psychology and Linguistics*, Springer, Netherlands, Dordrecht, 1987, pp. 85–95. ISBN 978-94-009-3645-4. doi:10.1007/978-94-009-3645-4_7.

[39] D. Marcu, *The Theory and Practice of Discourse Parsing and Summarization*, MIT Press, Cambridge, MA, USA, 2000. ISBN 0262133725.

[40] M.-F. Moens, Argumentation mining: How can a machine acquire common sense and world knowledge?, *Argument & Computation* **9**(1) (2017), 1–14.

[41] G. Morio, H. Ozaki, T. Morishita and K. Yanai, End-to-end argument mining with cross-corpora multi-task learning, *Transactions of the Association for Computational Linguistics* **10** (2022), 639–658. doi:10.1162/tacl_a_00481.

[42] S.B. Needleman and C.D. Wunsch, A general method applicable to the search for similarities in the amino acid sequence of two proteins, *Journal of Molecular Biology* **48**(3) (1970), 443–453, https://www.sciencedirect.com/science/article/pii/0022283670900574. doi:10.1016/0022-2836(70)90057-4.

[43] L.T. Nguyen, L. Van Ngo, K. Than and T.H. Nguyen, Employing the correspondence of relations and connectives to identify implicit discourse relations via label embeddings, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4201–4207, https://aclanthology.org/P19-1411. doi:10.18653/v1/P19-1411.

[44] A. Nie, E. Bennett and N. Goodman, DisSent: Learning sentence representations from explicit discourse relations, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 4497–4510, https://aclanthology.org/P19-1442. doi:10.18653/v1/P19-1442.

[45] J. Opitz, Argumentative relation classification as plausibility ranking, in: *Proceedings of the 15th Conference on Natural Language Processing, KONVENS 2019*, Erlangen, Germany, October 9–11, 2019, 2019, https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_51.pdf.

[46] J. Opitz and A. Frank, Dissecting content and context in argumentative relation analysis, in: *Proceedings of the 6th Workshop on Argument Mining*, Association for Computational Linguistics, Florence, Italy, 2019, pp. 25–34, https://aclanthology.org/W19-4503. doi:10.18653/v1/W19-4503.

[47] B. Pan, Y. Yang, Z. Zhao, Y. Zhuang, D. Cai and X. He, Discourse marker augmented network with reinforcement learning for natural language inference, in: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 989–999, https://aclanthology.org/P18-1091. doi:10.18653/v1/P18-1091.

[48] A. Panchenko, E. Ruppert, S. Faralli, S.P. Ponzetto and C. Biemann, Building a web-scale dependency-parsed corpus from CommonCrawl, in: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, European Language Resources Association (ELRA), Miyazaki, Japan, 2018, https://aclanthology.org/L18-1286.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, Scikit-learn: Machine learning in python, *Journal of Machine Learning Research* **12**(85) (2011), 2825–2830, http://jmlr.org/papers/v12/pedregosa11a.html.

[50] A. Peldszus and M. Stede, An annotated corpus of argumentative microtexts, in: *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, Vol. 2, 2015, pp. 801–815.

[51] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi and B. Webber, The penn discourse TreeBank 2.0, in: *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, European Language Resources Association (ELRA), Marrakech, Morocco, 2008, http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

[52] J.W.G. Putra, S. Teufel and T. Tokunaga, Multi-task and multi-corpora training strategies to enhance argumentative sentence linking performance, in: *Proceedings of the 8th Workshop on Argument Mining*, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 12–23, https://aclanthology.org/2021.argmining-1.2. doi:10.18653/v1/2021.argmining-1.2.

[53] L. Qin, Z. Zhang, H. Zhao, Z. Hu and E. Xing, Adversarial connective-exploiting networks for implicit discourse relation classification, in: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 1006–1017, https://aclanthology.org/P17-1093. doi:10.18653/v1/P17-1093.

[54] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li and P.J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *Journal of Machine Learning Research* **21**(140) (2020), 1–67, http://jmlr.org/papers/v21/20-074.html.

[55] N. Reimers and I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992, https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[56] G. Rocha and H.L. Cardoso, Context matters! Identifying argumentative relations in essays, in: *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing, SAC '22*, Association for Computing Machinery, New York, NY, USA, 2022, pp. 879–882. ISBN 9781450387132. doi:10.1145/3477314.3507246.

[57] G. Rocha, C. Stab, H. Lopes Cardoso and I. Gurevych, Cross-lingual argumentative relation identification: From English to Portuguese, in: *Proceedings of the 5th Workshop on Argument Mining*, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 144–154, https://aclanthology.org/W18-5217. doi:10.18653/v1/W18-5217.

[58] A. Rutherford and N. Xue, Improving the inference of implicit discourse relations via classifying explicit discourse connectives, in: *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Denver, Colorado, 2015, pp. 799–808, https://aclanthology.org/N15-1081. doi:10.3115/v1/N15-1081.

[59] C. Schulz, S. Eger, J. Daxenberger, T. Kahse and I. Gurevych, Multi-task learning for argumentation mining in low-resource settings, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 35–41, https://aclanthology.org/N18-2006. doi:10.18653/v1/N18-2006.

[60] W. Shi and V. Demberg, Learning to explicitate connectives with Seq2Seq network for implicit discourse relation classification, in: *Proceedings of the 13th International Conference on Computational Semantics – Long Papers*, Association for Computational Linguistics, Gothenburg, Sweden, 2019, pp. 188–199, https://aclanthology.org/W19-0416. doi:10.18653/v1/W19-0416.

[61] D. Sileo, T. Van De Cruys, C. Pradel and P. Muller, Mining discourse markers for unsupervised sentence representation learning, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3477–3486, https://aclanthology.org/N19-1351. doi:10.18653/v1/N19-1351.

[62] C. Stab and I. Gurevych, Identifying argumentative discourse structures in persuasive essays, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang and W. Daelemans, eds, Association for Computational Linguistics, Doha, Qatar, 2014, pp. 46–56, https://aclanthology.org/D14-1006. doi:10.3115/v1/D14-1006.

[63] C. Stab and I. Gurevych, Parsing argumentation structures in persuasive essays, *Computational Linguistics* **43**(3) (2017), 619–659, https://aclanthology.org/J17-3005. doi:10.1162/COLI_a_00295.

[64] S. Stevens-Guille, A. Maskharashvili, X. Li and M. White, Generating discourse connectives with pre-trained language models: Conditioning on discourse relations helps reconstruct the PDTB, in: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, O. Lemon, D. Hakkani-Tur, J.J. Li, A. Ashrafzadeh, D.H. Garcia, M. Alikhani, D. Vandyke and O. Dušek, eds, Association for Computational Linguistics, Edinburgh, UK, 2022, pp. 500–515, https://aclanthology.org/2022.sigdial-1.48. doi:10.18653/v1/2022.sigdial-1.48.

[65] S. Tafreshi, J. Sedoc, A. Rogers, A. Drozd, A. Rumshisky and A. Akula (eds), *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, Association for Computational Linguistics, Dublin, Ireland, 2022, https://aclanthology.org/2022.insights-1.0. doi:10.18653/v1/2022.insights-1.

[66] O. Toledo-Ronen, M. Orbach, Y. Bilu, A. Spector and N. Slonim, Multilingual argument mining: Datasets and analysis, in: *Findings of the Association for Computational Linguistics: EMNLP 2020*, Association for Computational Linguistics, 2020, pp. 303–317, Online, https://aclanthology.org/2020.findings-emnlp.29. doi:10.18653/v1/2020.findings-emnlp.29.

[67] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.U. Kaiser and I. Polosukhin, Attention is all you need, in: *Advances in Neural Information Processing Systems*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds, Vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[68] I. Vulić and N. Mrkšić, Specialising word vectors for lexical entailment, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1134–1145, https://aclanthology.org/N18-1103. doi:10.18653/v1/N18-1103.

[69] H. Wang, Z. Huang, Y. Dou and Y. Hong, in: *Argumentation Mining on Essays at Multi Scales, in: Proceedings of the 28th International Conference on Computational Linguistics*, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 5480–5493, https://aclanthology.org/2020.coling-main.478. doi:10.18653/v1/2020.coling-main.478.

[70] B. Webber, R. Prasad, A. Lee and A. Joshi, A discourse-annotated corpus of conjoined VPs, in: *Proceedings of the 10th Linguistic Annotation Workshop Held in Conjunction with ACL 2016 (LAW-X 2016)*, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 22–31, https://aclanthology.org/W16-1704. doi:10.18653/v1/W16-1704.

[71] Y. Ye and S. Teufel, End-to-end argument mining as biaffine dependency parsing, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Association for Computational Linguistics, 2021, pp. 669–678, Online, https://aclanthology.org/2021.eacl-main.55. doi:10.18653/v1/2021.eacl-main.55.